DELHI UNIVERSITY
LIBRARY

DELHI UNIVERSITY LIBRARY

Cl. No. MX411:B J6

Ac. No. 236202 Date of release for loan 1 APR 1968

This book should be returned on or before the date last stamped below. An overdue charge of **5 Paise** will be collected for each day the book is kept overtime.

METHUEN'S MONOGRAPHS
ON APPLIED PROBABILITY AND STATISTICS

General Editor: M. S. BARTLETT F.R.S.

THE THEORY OF ROAD TRAFFIC FLOW

The Theory of Road Traffic Flow

WINIFRED D. ASHTON

King's College, London

LONDON: METHUEN & CO LTD
NEW YORK: JOHN WILEY & SONS INC

First published 1966

© 1966 *Winifred D. Ashton*

Printed in Great Britain by
Spottiswoode, Ballantyne and Co Ltd
London and Colchester

Contents

Preface *page* vii

1 Introduction 1

2 The Fundamental Diagram of Traffic I 7

2.1 Definition and measurement of traffic stream parameters 7
2.2 The fundamental diagram 17

3 Dynamical Theories: Follow-the-Leader Models of Traffic Flow 21

3.1 Introduction 21
3.2 Mathematical equations of simple follow-the-leader models 23
3.3 Alternative derivations of the steady-state equations 33
3.4 Extensions of the stimulus-response equation 35

4 Kinematic Theories and Fluid Analogies 39

4.1 Introduction 39
4.2 The continuum theory 39
4.3 More complicated models 45
4.4 Connection with follow-the-leader theories 47
4.5 Boltzmann-like theory of traffic flow 48

5 The Fundamental Diagram of Traffic II 52

6 From Traffic Signals to Traffic Cybernetics 57

6.1 Introduction 57
6.2 Traffic signals 57
6.3 Formula for the delay at a signal-controlled intersection 59
6.4 Optimum settings for fixed-cycle signals 63
6.5 Pedestrian crossings 67
6.6 The future 70

7 Statistical Distributions and Queueing Theory 77
7.1 Random distributions 77
7.2 Pearson's distributions 87
7.3 Queueing theory 90
7.4 Application of queueing theory to overtaking 102

8 The Stochastic Approach to Static Delay Problems 106
8.1 Introduction 106
8.2 Simple models of static delay at priority intersections 110
8.3 More complicated models 120
8.4 Signal-controlled intersections 124
8.5 Delays on an open road 132
8.6 Bottlenecks 133

9 Simulation of Traffic Problems 135
9.1 Introduction 135
9.2 A queueing model for a T-junction 140
9.3 Sampling from a distribution 143

10 Accident Statistics 148
10.1 Accident proneness and other problems 148
10.2 Theoretical distributions 151
References 169

Subject Index 175

Author Index 177

Preface

Until recently accounts of work in the field of traffic flow have appeared only in articles scattered through a variety of technical journals relating to different scientific disciplines. Access to the literature even for the theoretician, let alone the practising traffic engineer, has, therefore, been none too easy. Recently there have been two attempts to provide more comprehensive accounts of the theory. One of these, *Mathematical Theories of Traffic Flow* (Academic Press, New York, 1963) by F. A. Haight, is largely concerned with the stochastic approach. The other, *An Introduction to Traffic Flow Theory* edited by D. L. Gerlough and D. G. Capelle, is a paperback publication by the Highway Research Board of America. It consists of a collection of articles on different aspects of traffic flow theory by a number of different authors, and is in the nature of an interim survey.

This book is the first comprehensive textbook to be published in this country. It is expected that it will be of interest to statisticians, applied mathematicians and engineers specializing in the traffic field. Although the approach is theoretical, the needs of the practising engineer have not been forgotten, and it is hoped that a number of the results collected here will find practical application in the field.

As far as mathematical prerequisites are concerned, a general mathematical background including some elementary knowledge of statistical distributions and queueing theory is desirable. A certain amount of the necessary material is, however, included in the text.

Acknowledgement is made to the editors or honorary editors of the following journals for permission to use material published therein: *Biometrika*, *British Computer Journal*, *Journal of the American Statistical Association*, *Operations Research*, *Proceedings of the Institution of Civil Engineers*, *Siam Review* and *Traffic Engineering and Control*. Acknowledgement is also made to the authors of the papers concerned, in particular to J. C. Butcher, L. C. Edie, G. F. Newell, M. S. Raff, J. C. Tanner, J. G. Wardrop, F. V. Webster and P. I. Welding.

PREFACE

The author would like to thank Dr F. Garwood, Head of Statistics, and Planning, Traffic and Safety Division, Road Research Laboratory for access to various publications, and discussions about the pedestrian control experiment described in §6.5. Thanks are also due to the Ministry of Transport for permission to describe this experiment. Acknowledgement is also made to Mr A. G. Hawkes, University College, London for comments and suggestions. The author acknowledges with particular gratitude the help of Professor D. R. Cox, Birkbeck College, London for reading the manuscript and for general encouragement in completing the project.

WINIFRED D. ASHTON

King's College,
London
March, 1965.

CHAPTER 1

Introduction

In the past few years traffic problems have begun to present not only a technological but also a sociological challenge. The problem of the roads is part of the general transport problem which has arisen during the years since the Second World War. There has been a general rise in standards of living in every country in the world, and this has been accompanied by changes in social habits. In Britain there has been a considerable growth in the use of personal transport, and similar increases are evident in many countries. Percentage increases in vehicle registrations over the years 1953–1963 vary considerably from country to country due to several factors, including the number of vehicles already on the road at the beginning of the period. But everywhere the trend is the same. It has been estimated that ten and a half million vehicles were on the roads of Britain in 1963, and this figure is expected to increase to about eighteen million by 1970.

Of the total number of vehicles on the road, roughly five out of every six are private cars. These are no longer the prerogative of the privileged few, and as a consequence, the present-day population is more mobile than it has ever been. But the era of the private car has brought not only benefits but problems of various kinds, for instance traffic jams, parking difficulties and increased accident rates. The question arises, therefore, of how to organize road traffic so that the benefits of the increased mobility can be enjoyed to the full.

Increasing interest is now being shown both by practising traffic engineers and by theoreticians of various kinds, mainly mathematicians and statisticians. The former have begun to feel that rule of thumb methods must be supplemented or superseded by theoretical analysis. The latter have found that traffic flow problems constitute a stimulating intellectual challenge.

Problems of the road are complex. They involve human nature on the one hand and physical laws of time, space and movement on the other. They are also many-sided: they involve constructional, legal, educational and administrative factors. The road system is inadequate to cope with the demands now being made upon it and the result is to be found in accidents and economic waste due, for example, to delays from congestion. In several large cities in Britain the average speed of traffic in city-centres is 10 m.p.h. or less. London Transport have stated that if their buses were able to travel only one mile an hour faster it would result in a saving of £1,000,000 per annum! The job of the traffic engineer is twofold. He must fit the traffic to the existing road system by means of regulation and control, and he must fit the road system to the traffic by planning and design. His objective is to achieve maximum speed and capacity with safety.

In his report *Traffic in Towns*, published in 1963 at the end of 3 years research, Professor Colin Buchanan has taken a grim and searching look into the future. He suggests a revolutionary solution to the nightmare of traffic congestion that now faces us in this country. He also gives a blunt warning that there is no simple solution to living with the motor-vehicle, and that unless steps are taken it will defeat its own utility and bring about a disastrous degradation in the surroundings for living. In Buchanan's words we must meet the challenge 'without confusion over purpose, without timidity over means, and above all without delay'.

Among others, the Buchanan report makes the following three recommendations. First, the multi-level reshaping of cities so that vehicles and pedestrians can be separated. Two or three tiers are envisaged, of which the lowest, below the present ground level, would take the main roads. At ground level would be roads for local traffic with parking space underneath the shops. Above this, on a raised platform, would be pedestrian precincts with entries to shops and moving stairways. Second, the creation of living areas in which through traffic would be banned, but which would be linked to each other by a network of roads. Third, a restriction of some sort on the number of vehicles to be allowed in congested city-centres. Pricing the use of road space or parking restrictions would be two alternatives here. These measures would have to be linked with the provision of adequate and cheap public transport to give an acceptable alternative for commuter travel. The report emphasizes that buildings and roads must not be considered separately, but merged in the new theme of 'traffic architecture'.

Every local authority in Britain is to reconsider its development and road program in the light of this report. Annual spending on urban roads is to be stepped up from £50 million in 1963 to £140 million by 1970. Unlimited funds, however, will never be available and there is a pressing need for research into traffic problems and priorities.

Real traffic situations are usually very complex. In theoretical studies using mathematical models we must usually choose simple problems, leaving the analysis of complicated traffic networks to simulation on a computer. It is necessary, however, to combine empirical and theoretical studies: theory can suggest new approaches, whilst experiment and observation can suggest modifications to the theory which make it more realistic. This is the essence of the scientific method: a theory is tested qualitatively and, if not adequate, altered to fit in with measurements made in the field.

At least five approaches to traffic problems are possible. The earliest approaches were empirical and intuitive. Tables of such things as accident statistics have been compiled for some years. Out of this simple approach comes the attempt to determine the connections between various relevant quantities which have been measured in the field. Analysis of data by known statistical techniques has also been carried out. As early as 1935 Greenshields made an experimental study of traffic flow by measuring actual flows and velocities for one lane of traffic, and fitted a straight line to a plot of velocity against concentration. Time series analysis has also been found helpful in examining problems connected with the disruption in the regularity of bus services due to traffic congestion. In addition, the application of standard statistical techniques in 'before-and-after-studies' has helped to avoid the uneconomical use of large samples.

The general field of instrumentation is continually being improved, and this is specifically true of instruments furnishing data in a form suitable for input to an electronic computer. The analysis of traffic surveys is now carried out by such machines, and the results enable traffic engineers to forecast future traffic volumes and to assign the traffic to a network of roads. Thus computers can help in traffic-planning projects. Information from origin-destination surveys, together with forecasts of future traffic volumes lead to the design of networks to deal with this volume. Detailed studies have been made, not only of the capacities of roads, but also of the demand for travel within and between cities. It is now possible to analyse the traffic census before the information in it becomes completely out of date. Thus a realistic estimate of future traffic volumes can be obtained, and the merits of alternative traffic and road schemes evaluated.

Of course the amassing and analysis of statistical data is not an end in itself. This basically empirical approach has not provided satisfactory answers to many of the most important problems posed about road traffic. Why, for example, does dense traffic move in a series of stops and starts? In order to answer such a question some aspect of the problem must be isolated and then a mathematical model set up. This method of attack belongs to what is now called Operational or Operations Research.

One possible theoretical approach is through traffic dynamics, otherwise known as car-following or follow-the-leader theory. This is concerned with the dynamics of a single lane of dense traffic, the vehicles following one another without overtaking. Much about this sort of flow is not yet completely understood because its dynamics of stop and start are not simple. When the stream is stopped at a traffic signal showing red the vehicles are packed bumper to bumper. When the signal changes to green the stream does not start as a complete unit: the first vehicle starts, then the next and so on. In other words a wave front passes down the line of vehicles. Observation has indicated that the velocity of this starting wave is fairly constant at about 25 m.p.h., a figure which is dependent on both the reaction-time of drivers and on the mechanical properties of their motor-vehicles. This problem is, of course, a very small one compared with the whole subject of traffic but a thorough understanding of its behaviour would be very useful. This theory is developed in detail in Chapter 3.

Hydrodynamic analogies and kinematic wave theory have provided another method of attack. Certain results about the propagation of 'waves' and 'shock-waves', regarding traffic as a cooperative phenomenon, have been obtained. This theory does not take account of the properties of individual vehicles, so that it cannot be expected to explain traffic behaviour in detail; also it applies to traffic of low to medium concentrations only. This approach, originally due to Lighthill and Whitham (1955), is described in Chapter 4.

The dynamic and kinematic theories just outlined are concerned with deterministic relationships. Such relationships are adequate for describing traffic behaviour in a macroscopic form, and can provide a sound basis for general calculations in traffic engineering. More detailed descriptions on a microscopic basis involve probabilistic flow models and queueing theory. The complexity of the statistical theory involved has tended to inhibit development in this direction, and only simple situations can be dealt with in this way at the present time. These stochastic models deal both with junctions, where the relevant quantity of interest is either the average delay to vehicles or the capacity of the junction, and also with lengths of road, where the relevant quantity is the average speed of a journey. Some problems of this sort are dealt with in Chapter 8.

Both deterministic and probabilistic models are usually conceived in mathematical terms, but an enormous quantity of numerical and graphical data has been accumulated over the past half century, and this must be used in conjunction with the theoretical models. It must be realized that the value of a contribution to road traffic theory depends on its potential field of application to the real situation, although sometimes a theoretical analysis may give some insight into the nature of the situation without leading at once to a specific practical application. Theoretical work, of course, eventually becomes the stock-in-trade of the practising engineer, and although rule-of-thumb methods may give results, the greatest degree of efficiency can only be achieved when due attention is paid to theory.

A quite different method for tackling traffic problems is to use simulation on an electronic computer. In other words we attempt to represent reality by means of machine techniques. This is usually only resorted to when analytical methods fail, but traffic networks of large cities and the more complicated traffic situations are likely to be amenable only to this approach.

The question of simulation leads to a consideration of the part played by computers in traffic engineering. They were used first as substitutes for traditional manual computation in such things as traffic accident analysis, studies of traffic movement and volume patterns, and in the statistical analysis of experimental data. But the real potentialities of the computer in the future lie in less traditional applications such as simulation studies. This aspect is dealt with in Chapter 9.

Until quite recently – in fact until 1959, the date of the first International Symposium on Traffic Flow – the expression 'traffic flow theory' did not mean anything very definite. Perhaps the most important job of this first Symposium was to define the field: it lies somewhere between traffic engineering, applied mathematics and operational research. The second International Symposium, held in London in 1963, brought together the existing body of theory and results, and the third, held in New York in 1965, extended them further. The *Proceedings* of the first two Symposia (1961, 1965) are now available.

The various theories of traffic flow outlined above are discussed in detail in Chapters 2 to 9. Chapter 10 deals with the related topic of road accidents.

CHAPTER 2

The Fundamental Diagram of Traffic I

2.1 Definition and Measurement of Traffic Stream Parameters

Before discussing the various theories of traffic flow we must define some of the terms which are used. In studying road traffic as a stream or continuum we are primarily interested in three characteristics of the stream, viz. the flow q, measuring quantity per unit time, the concentration (or density) k, measuring quantity per unit distance and the speed u, measuring distance per unit time. Associated with the idea of concentration or density is that of the gap or headway between two vehicles. The headway can be thought of in terms of either time or distance, measurement usually being made from front bumper to front bumper. The time headway is the one relevant to problems concerning pedestrians crossing a road or vehicles crossing the main stream at a T-junction. Note that the space headway is sometimes referred to in the literature as spacing and the term headway reserved for time-gaps; in what follows ambiguity has been avoided by using the adjectives time and distance wherever necessary.

Since road traffic is not in fact a continuous fluid but is made up of discrete vehicles, methods of observation are concerned with measurements on individual vehicles. The need to convert discrete measurements of, for example, time and space headways or vehicle speeds, into continuum characteristics has led to precise definitions of the latter in terms of the former.

The quantities q, k and u are meaningful only as averages (or probability distributions) and the kinds of averages to be used vary. In some cases it is correct to use arithmetic averages, in others harmonic averages; in some cases space averages are suitable, in others time averages. It is necessary, therefore, for the correct type of average for each situation to be developed from the definitions of flow, concentration and speed. These latter are themselves derived from two types of measurements, one made at a point in space and the other at a point in time.

The flow of traffic on a road can be described in two quite different ways. We can stand by the side of the road and count the number of vehicles passing per unit time (one minute or one hour, say). This gives the flow q. In general this is a function of the time (t) and of the position along the road (x), so that $q=q(t,x)$. Otherwise we can take an aerial photograph and note the numbers of vehicles at one and the same instant on unit length of road (one mile, say). This gives the concentration as $k=k(t,x)$.

The two quantities q and k are related to one another, and the mean speed of vehicles is $u=q/k$ since unit length of road containing k vehicles is emptied in time k/q. This flow-concentration (q-k) relationship describes how the two distributions in space and in time are related.

TABLE 2.1

Speed range (m.p.h.)	Mid-point u_i	Flow v.p.h. q_i	$\frac{100q_i}{Q}$	Concentration v.p.m. k_i	$\frac{100k_i}{K}$
2–5	3·5	1	0·3	0·3	1·7
6–9	7·5	4	0·8	0·5	3·3
10–13	11·5	0	0·0	0·0	0·0
14–17	15·5	7	1·5	0·4	2·9
18–21	19·5	20	4·5	1·0	6·9
22–25	23·5	44	9·9	1·9	12·7
26–29	27·5	80	17·9	2·9	19·7
30–33	31·5	82	18·2	2·6	17·7
34–37	35·5	79	17·5	2·2	14·9
38–41	39·5	49	10·9	1·2	8·3
42–45	43·5	36	7·9	0·8	5·4
46–49	47·5	26	5·7	0·6	3·8
50–53	51·5	9	1·9	0·2	1·1
54–57	55·5	10	2·2	0·2	1·2
58–61	59·5	3	0·8	0·1	0·4
Total		450	100·0	14·9	100·0

N.B. The speed-ranges are recorded to the nearest m.p.h., and hence they are actually 1½–5½, 5½–9½ etc.

Consider first the distribution of speed in time. If there are n subsidiary streams with flows $q_1, \ldots, q_n$ and velocities $u_1, \ldots, u_n$, the total flow is given by

$$Q = \sum_{i=1}^{n} q_i.$$

For the distribution of speed in space, the average time-interval between vehicles is $1/q_i$, and the distance travelled in this time is u_i/q_i. Then $k_i = q_i/u_i$ and

$$K = \sum_{i=1}^{n} k_i.$$

Wardrop (1952) has given an empirical table of q and k plotted against u: this data is reproduced in Table 2.1. If the speed is plotted against percentage frequencies in time and in space, as in Fig. 2.1 (a) and (b), it can be seen that the bulk of the time distribution is further to the right than that of the space distribution. This means that speeds are greater for the time distribution.

If we consider Table 2.1 we see that there is not just one mean speed, but two, viz. the time mean speed $\bar{u}_t$ and the space mean speed $\bar{u}_s$. The time mean speed is defined as the arithmetic mean of the speeds of vehicles passing a point during a given interval of time. The space mean speed is defined as the arithmetic mean of the speeds of vehicles occupying a given length of lane. So that

$$\bar{u}_t = \sum_{i=1}^{n} \frac{q_i u_i}{Q},$$

$$\bar{u}_s = \sum_{i=1}^{n} \frac{k_i u_i}{K}.$$

The latter quantity $\bar{u}_s = \sum_{i=1}^{n} q_i/K$, since $u_i = q_i/k_i$. It follows therefore, that

$$Q = K\bar{u}_s.$$

This means that it is the space mean speed that is involved in the flow-concentration relation: there is no equivalent relation for the time mean speed, $\bar{u}_t$.

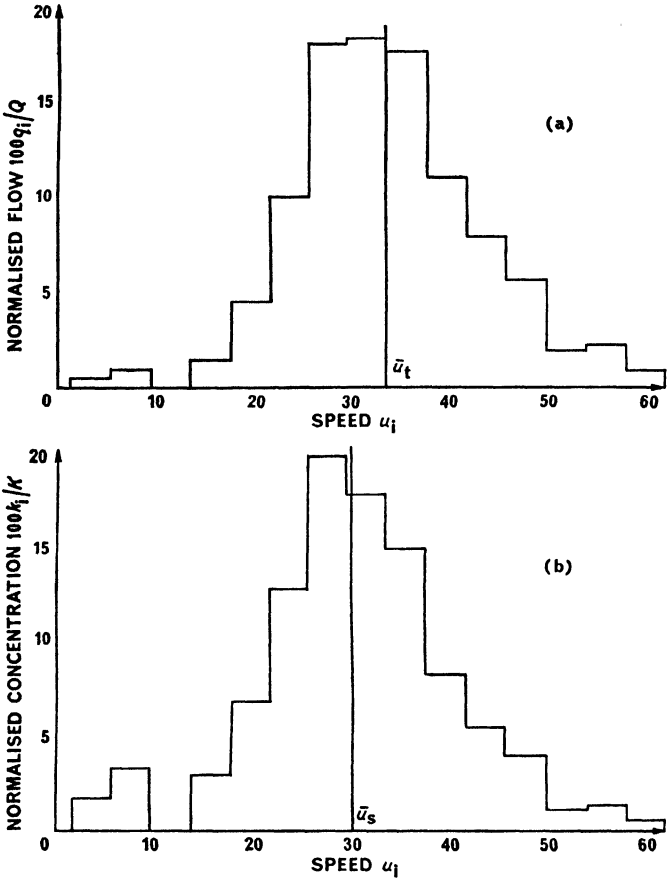

FIG. 2.1 (a) and (b). Plots of flow and concentration versus speed (empirical data).

For the data of Table 2.1 we have

$$Q = \sum_{i=1}^{n} u_i k_i = \sum_{i=1}^{n} q_i = 450,$$

giving $$\bar{u}_s = \frac{Q}{K} = \frac{450}{14{\cdot}9} \approx 30 \text{ m.p.h.}$$

and $$\bar{u}_t = \sum_{i=1}^{n} \frac{u_i q_i}{Q} = \frac{1508{\cdot}7}{450} \approx 33{\cdot}5 \text{ m.p.h.}$$

so that $\bar{u}_t > \bar{u}_s$.

Wardrop has in fact shown that

$$\bar{u}_t = \bar{u}_s + \sigma_s^2/\bar{u}_s,$$

where the variance

$$\sigma_s^2 = \sum_{i=1}^{n} k_i(u_i - \bar{u}_s)^2/K.$$

In practice $\bar{u}_t$ usually exceeds $\bar{u}_s$ by about 6–12%, except of course for the case of constant speed. Wardrop's formula is proved straightforwardly from the definitions of $\bar{u}_t$, $\bar{u}_s$ and σ_s^2 as follows:

$$\bar{u}_t = \frac{\sum_{i=1}^{n} q_i u_i}{Q} = \frac{\sum_{i=1}^{n} k_i u_i^2}{Q} = K\frac{\sum_{i=1}^{n} f_i u_i^2}{Q}, \text{ where } f_i = \frac{k_i}{K}.$$

Hence
$$\bar{u}_t = \frac{\sum_{i=1}^{n} f_i u_i^2}{\bar{u}_s} = \frac{\sum_{i=1}^{n} f_i[\bar{u}_s + (u_i - \bar{u}_s)]^2}{\bar{u}_s}$$

$$= \frac{\left[\sum_{i=1}^{n} f_i \bar{u}_s^2 + \sum_{i=1}^{n} f_i(u_i - \bar{u}_s)^2\right]}{\bar{u}_s}$$

since
$$\sum_{i=1}^{n} f_i(u_i - \bar{u}_s)\,\bar{u}_s = 0.$$

Whence
$$\bar{u}_t = \bar{u}_s + \sigma_s^2/\bar{u}_s.$$

Lighthill and Whitham (1955) have supplied precise definitions of the flow q and the concentration k at a point X and at a time t by giving instructions for measuring them over a short increment δx, whose midpoint is at X, and for a long period T, whose midpoint is at t, as follows:

$$q = n/T,$$

$$k = \frac{\sum \delta t_i}{T\,\delta x} = \frac{\sum \delta t_i/T}{\delta x},$$

where n = number of vehicles crossing δx in T,

δt_i = time taken by ith vehicle to cross δx,

and $\sum$ indicates summation from 1 to n.

Then
$$k = \frac{\text{average number of vehicles on } \delta x}{\text{length } \delta x}$$

$$= \text{number of vehicles/unit length of road.}$$

The speed of the stream is then defined as

$$\bar{u}_s = \frac{q}{k} = \frac{n\,\delta x}{\sum \delta t_i}.$$

This is the space mean speed of Wardrop: it is weighted according to the time spent on δx. From this it can be seen that

$$\bar{u}_s = \frac{n}{\sum 1/u_i} = \left[\frac{1}{n}\sum \frac{1}{u_i}\right]^{-1}, \qquad (2.1)$$

which means that $\bar{u}_s$ is the harmonic mean of the observed speeds, as measured for example by a radar speedmeter.

Wardrop also showed how to use two aerial photographs of a long section of roadway X taken a short time δt apart to obtain

$$q = \frac{\sum \delta x_i}{X\,\delta t},$$

$$k = \frac{n}{X},$$

$$\bar{u}_t = \frac{\sum \delta x_i}{n\,\delta t} = \frac{\sum u_i}{n}, \qquad (2.2)$$

where n = number of vehicles on the road section at time t,

and δx_i = distance travelled by the ith vehicle in δt.

In practice there are four main ways of obtaining the stream velocity:

(i) Successive vehicles (or every nth vehicle) are timed over a measured section X; $u_i = X/t_i$ is calculated for each, and then $\bar{u}_t$ is obtained from the formula

$$\bar{u}_t = \frac{\sum u_i}{n}.$$

It is also possible to obtain $\bar{u}_s$ from the formula

$$\bar{u}_s = \frac{n}{\sum 1/u_i}.$$

(ii) A radar speedmeter gives the u_i directly (these are known as spot speeds) and then $\bar{u}_t$ and $\bar{u}_s$ are found as in (i).

(iii) As (i) but $\bar{t}=\sum t_i/n$, is obtained first, and then $\bar{u}_s$ from

$$\bar{u}_s = \frac{X}{\bar{t}}.$$

(iv) Two aerial photographs taken at an interval δt give the distance δx_i covered by each vehicle, and then the speeds of individual vehicles are given by

$$u_i = \frac{\delta x_i}{\delta t},$$

so that

$$\bar{u}_t = \frac{\sum u_i}{n}.$$

A knowledge of the mean stream speeds is essential both in economic studies, and in before-and-after studies where the average effect on all traffic is being assessed. Which of the two possible values $\bar{u}_s$ or $\bar{u}_t$ is relevant depends on the particular situation. The relationships between vehicles in space and in time have a particularly important role to play in empirical studies since the time measurements, using road tapes and photoelectric cells, are easy for the traffic engineer to make. Time averages are usually more useful since journey-time rather than average speed is the important factor. On the other hand, journey-speed distributions tend to be more consistent and symmetrical than journey-time distributions, which are skew to the right.

We can express equations (2.1) and (2.2) together with their respective equations for q and k in terms of the speed u_i and time-headway h_i or spacing s_i of each vehicle, instead of in terms of δx and δt, as follows:

Space means	Time means
$q = \frac{n}{\sum h_i}$	$q = \frac{\sum u_i}{\sum s_i}$
$k = \frac{\sum 1/u_i}{\sum h_i}$	$k = \frac{n}{\sum s_i}$
$u = \frac{n}{\sum 1/u_i}$	$u = \frac{\sum u_i}{n}$

From these result it is clear that the stream speed is the harmonic mean of vehicle speeds measured at a point, and also the arithmetic mean of vehicle speeds measured at an instant. The flow q at a point is the reciprocal of the arithmetic mean of the time headways, and the concentration k is the reciprocal of the arithmetic mean of the spacing.

Two questions now arise. First, how should k be averaged over time at a point, or over space at an instant. Secondly, if the averages of q_i, k_i and u_i for groups of N_i ($i=1\dots n$) vehicles passing a point have been computed, what averages are to be used for combining two or more of the groups. Edie (1963) has developed new definitions independent of the method of measurement, which unify the definitions given above and eliminate any ambiguity about the method of averaging to be used. Previously investigators have taken their observations over a short stretch of road or a short period of time, but it has now become desirable to study traffic behaviour over longer sections of road and different periods of time. For this purpose it is necessary to have definitions applicable to vehicle trajectories in any space-time domain.

If $t_i=N_i/q_i$ is the duration of the ith flow level, $\sum\limits_{i=1}^{n} N_i=N$ and $\sum\limits_{i=1}^{n} t_i=T$, the short roadway observations combined for different times are as follows:

$$q=\frac{N}{T}=\frac{\sum t_i q_i}{\sum t_i}, \tag{2.3}$$

$$k=\frac{\sum 1/u_i}{T}=\frac{\sum t_i k_i}{\sum t_i}, \tag{2.4}$$

$$u=\frac{N}{\sum 1/u_i}=\frac{\sum t_i q_i}{\sum t_i k_i}.$$

Equations (2.3) and (2.4) give the time-weighted averages of flow and concentration and their quotient gives the average speed.

If $x_i=N_i/k_i$ is the distance over which the concentration k_i is measured, the short time observations combined for different road sections are as follows:

$$q=\frac{\sum u_i}{\sum s_i}=\frac{\sum x_i q_i}{\sum x_i}, \tag{2.5}$$

$$k = \frac{N}{X} = \frac{\sum x_i k_i}{\sum x_i}, \tag{2.6}$$

$$u = \frac{\sum u_i}{N} = \frac{\sum x_i q_i}{\sum x_i k_i}.$$

Equations (2.5) and (2.6) give the space-weighted averages of flow and concentration and their quotient gives the average speed.

Multiplying the numerator and denominator of equation (2.3) by δx gives

$$q = \frac{N\,\delta x}{T\,\delta x} = \frac{\text{Total distance travelled by } N \text{ vehicles on } \delta x}{\text{Area of rectangular space-time domain } (\delta x . T)}.$$

Equation (2.4) gives

$$k = \frac{\sum \delta t_i}{T\,\delta x} = \frac{\text{Total time spent on } \delta x \text{ by } N \text{ vehicles}}{\text{Area of space-time domain } (\delta x . T)}.$$

Equation (2.5) gives

$$q = \frac{\sum \delta x_i}{X\,\delta t} = \frac{\text{Total distance travelled by } N \text{ vehicles on } \delta x}{\text{Area of space-time domain } (\delta t . X)}.$$

Equation (2.6), multiplying by δt gives

$$k = \frac{N\,\delta t}{X\,\delta t} = \frac{\text{Total time spent on } \delta x \text{ by } N \text{ vehicles}}{\text{Area of space-time domain } (\delta t . X)}.$$

Then the original two sets of definitions can be combined into a single set which is independent of the methods of measurement. Putting the space-time domain $(\delta x . T) = (\delta t . X) = A$

$$q = \frac{\sum \text{distances travelled in } A}{A},$$

$$k = \frac{\sum \text{time spent in } A}{A},$$

and then
$$u = \frac{\sum \text{distances travelled}}{\sum \text{times spent}}.$$

In order to combine n values of q_i, k_i and u_i for domains with areas a_i the following formulae are used:

$$q = \frac{\sum a_i q_i}{\sum a_i}, \quad k = \frac{\sum a_i k_i}{\sum a_i}, \quad u = \frac{\sum a_i q_i}{\sum a_i k_i}.$$

These definitions of q, k and u reduce the sensitivity to random errors which are always incurred in measurements at a point or for an instant. Errors of as much as 20% can be observed in averages taken over a number of vehicles or periods of time for observations at a point. Many of these errors arise from the wrong use of arithmetic means for q, k and u where harmonic means are appropriate. The use of the arithmetic mean for all these quantities always results in an overestimate. Hence, if q is obtained from the product of u and k a very large error in q may result.

Another traffic parameter which has been suggested (Jones and Potts, 1962) as being useful for comparing different traffic situations is the acceleration dispersion σ. This arises in the following manner. On an open road a driver usually attempts to maintain a constant speed. Because of road conditions and other factors, this ideal state is not in general achieved, and in reality the movement of the vehicle follows an acceleration pattern which is a function of the time. This pattern can be described by an acceleration distribution function which is roughly Normal. The random component of this is called 'acceleration noise', and in general it has both a longitudinal and a transverse component. The latter could be large for, say, a winding road, but as yet this component has not been much studied. Considering the longitudinal component only, the smoothness or otherwise of the driving can be measured by the dispersion parameter σ defined by

$$\sigma^2 = \frac{1}{T}\int_0^T \left[\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}\right]^2 \mathrm{d}t$$

where T = actual running time,

and $\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}$ = acceleration along the road at time t.

The dimensions of σ are those of acceleration and are measured in ft/sec^2 or as a fraction of g. The value of σ varies according to road conditions and may be as much as 2 ft/sec^2.

If the acceleration is sampled at intervals of time Δt, then σ^2 is given by

$$\sigma^2 = \frac{1}{T}\sum \left[\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}\right]^2 \Delta t,$$

and it can be measured by using the approximation

$$\sigma^2 \approx \frac{1}{T}\sum\left[\frac{\Delta u}{\Delta t}\right]^2 \Delta t$$

$$= \frac{(\Delta u)^2}{T}\sum\frac{n^2}{\Delta t}$$

where Δt is the time taken for a change $n\,\Delta u$ (n integral) in the speed, and Δu is taken constant over the road.

Apparatus for making an automatic record of the acceleration noise, and for converting it into a form which is suitable for direct computer input, has already been developed in the United States.

The parameter σ is useful in stability analysis for rural and suburban roads. However, it is not always sensitive to changing traffic conditions in a congested environment where traffic moves with a low average speed. It fails to distinguish clearly between journeys at low speeds in dense traffic and faster journeys with the same acceleration noise mainly caused by traffic signals.

It has been suggested (Helly and Baker, 1965) that a modified parameter, G, be used instead. This is defined by

$$G = \frac{\sigma}{\bar{v}}$$

where $\bar{v}$ is the average speed. Hence G is a measure of the mean change in speed per unit distance of the journey.

This parameter conforms better than σ to subjective evaluations. Values of G in the range (0·0–0·02) are associated with a comfortable journey, values in the range (0·02–0·10) are acceptable, values in the range (0·10–0·20) cause frustration and values greater than 0·20 are unendurable.

2.2 The Fundamental Diagram

The relationship between q, the flow of traffic in vehicles per hour, and k, the concentration in vehicles per mile, has been called the fundamental diagram for traffic. It has engaged the interest of traffic engineers for many years, and its importance can be illustrated by a quite simple argument. Since the flow must be zero when the concentration is zero then, if it is assumed that the flow falls to zero at a concentration k_j (jam concentration), for values of the concentration between these limits the flow must rise to at least one maximum and the shape of the curve must be roughly as in Fig. 2.2.

Experimental data seem to confirm this argument to some extent, in that they show that as the concentration rises the flow first of all increases and reaches a maximum and then apparently falls, but there is some conflicting evidence regarding the exact shape of the curve. Note that k has a theoretical maximum corresponding to jam concentration k_j, but although u and q may have maxima, it is impractical to assign absolute upper limits to flow or speed. Of course $q(k) = k \, . \, u(k)$ is a relationship between real functions of k, and in a particular situation there may be an observable maximum value of q as a function of k. This maximum is called the 'capacity' of the road.

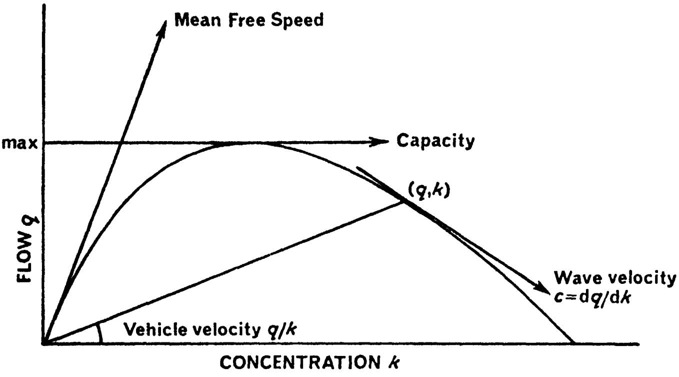

FIG. 2.2. The fundamental diagram or flow-concentration curve.

The capacity of a road is a quantity of the greatest interest to the traffic engineer, being relevant both in the design of new roads and in the efficient use of existing ones. It is, in fact, a concept which has been extensively investigated, but without general agreement on the exact definition of the concept or on methods of measuring it. However one of the most important results of the Lighthill–Whitham theory of kinematic waves (described in Chapter 4) has been to establish one method of measuring the capacity in practice. There are three different aspects of this quantity. 'Possible capacity' is usually defined as the maximum number of vehicles which can pass a given point in one hour under the prevailing roadway and traffic conditions.

'Practical capacity' is defined as the maximum number of vehicles which can pass a given point in one hour under the prevailing conditions without unreasonable delay or restriction in the drivers' freedom to manoeuvre. 'Design capacity' is similar to practical capacity as far as numerical values are concerned.

Let us now examine the proposed flow-concentration curve critically. The usual argument for the q-k curve involves two basic assumptions; that there is a single relationship between q and k for all possible values of k, and that the flow drops to zero at jam concentration, k_j. And it has been seen that these give rise to a curve which rises to a maximum and then falls again to zero.

If we attempt to look at the situation critically and from first principles the things which can be put forward are rather different. In the first place, if there were no interactions the flow would be u_f, where u_f is an average 'free speed'. The real flow at any concentration is likely to be less than this, and so the q-k curve will rise from zero with a slope which gradually decreases from its original value of u_f. Next, one of the parameters of the problem is the average effective vehicle length (usually taken as 15 ft in Britain), and this might be regarded as the reciprocal of a 'jam concentration', but the idea that the flow should fall smoothly to zero at this spacing should be questioned. A flow system in which an increase in concentration results in a decrease in flow seems at first sight curious; it would hardly be surprising if it proved to be unstable, and this is confirmed by the analysis of §3.4. Finally it is unlikely that the q-k relationship is the same for all values of k. In particular there may be one law for low concentrations and another for high concentrations. The arguments for this are as follows.

The interactions which occur between vehicles in a stream of traffic are very complicated. At low concentrations a given driver probably only takes account of the vehicle ahead, and then only to a small extent. As the concentration is increased, several factors begin to influence him. His manoevring distance is decreased, then the finite size of the vehicle ahead obscures his view of the road, until finally he becomes aware of the finite size of his own vehicle. At some stage he may also become conscious of the vehicles further ahead of and/or behind him. These are all factors which may affect the flow-concentration relationship for steady uniform flow.

Further effects occur when steady non-uniform flow is considered.

In this case is seems reasonable to suppose that a driver's behaviour will be different if he sees that he is moving into a region of heavy traffic from his behaviour when he sees that the traffic ahead is less dense than that near him. Finally, in unsteady flow, a driver is likely to be influenced quite strongly by local fluctuations in concentration.

A discontinuity near the maximum has been remarked on by Edie, who suggested a second law for the high density region.

Further discussion of the fundamental diagram and its derivation from the various theories of traffic flow will be found in Chapter 5.

CHAPTER 3

Dynamical Theories; Follow-the-leader Models of Traffic Flow

3.1 Introduction

There are a number of physical phenomena in nature which arise because small disturbances become amplified. One such example is turbulence in fluids, and traffic flow can be affected in a similar way. This phenomenon is usually described as instability.

There are many effects which occur in traffic as a result of instability. At the end of a long queue behind a bottleneck movement is of the stop-and-go variety, although the flow through the bottleneck may be quite steady.

The importance of both experimental and theoretical studies of instability has been recognized only recently. For many years experiments were directed towards producing empirical relations between flow and concentration, velocity and concentration and so on, under conditions of smooth flow. Theories such as car-following, which were designed to describe traffic flow at high concentrations were also, until recently, based upon the assumption that drivers behave in a stable manner. A potential instability was then introduced by bringing in a 'time-lag' factor. This time-lag in the response of a following vehicle to the accelerations and decelerations of the one immediately ahead causes a small initial disturbance which can then grow as it propagates down a long line of vehicles until finally a large disturbance has been generated from very little. This theory is developed mathematically in §3.2.

The first time-lag theories were linear theories designed to describe how a small disturbance is propagated along a chain of vehicles. They were capable of defining a range of the time-lag T for which the motion was stable, and of describing the propagation of waves under such conditions. They also gave the complementary range of T for which the motion was unstable, but the assumption that the initial disturbance was of small amplitude made it impossible to follow its development after it had become large.

Subsequently these theories were generalized, by the inclusion of a time-lag in a non-linear theory. This is also described in §3.2. The possibility of such a theory succeeding in describing all the observed phenomena has not yet been ruled out, but it does present some difficulties. The only time-constants in the model – for example T – are of the order of a few seconds. The question arises of how to explain an oscillation with a period of several minutes from a theory which has times measured in seconds. Extensive experiments conducted by the New York Port Authority in the Holland and Lincoln tunnels in New York City, where vehicles are not allowed to change lanes and concentrations are high for long periods at a time – in other words where car-following theories might be expected to apply – have presented a baffling phenomenon. It was noticed that when the concentration reached a certain intensity the steady flow of traffic broke down. About every 4 to 6 minutes a disturbance originated towards the exit of the tunnel, and as the disturbance propagated from one vehicle to the next, it became amplified until some vehicle came to a complete stop. This stopping wave then propagated back towards the entrance of the tunnel but the duration of the stoppage increased as it travelled. The deceleration wave travelled back faster than the subsequent acceleration wave so that by the time the stoppage rached the entrance about a mile back, entering vehicles might be stopped for as long as a minute. The flow through the bottleneck and beyond it, however, remained fairly steady.

The car-following or continuum theories described in §3.2 and Chapter 4 respectively have, however, succeeded in describing some traffic phenomena, at least qualitatively. For example, they have explained the existence of shocks and expanding waves, and the relation between the wave velocities and the flow-concentration or velocity-headway curves. They have also shown how certain average flow characteristics are related to the behaviour of individual vehicles.

The theory described in §3.4 contains the most important features of these previous theories, but includes also some modifications that admit of the description of certain other phenomena (particularly the instability) which were incompatible with previous theories. The structure of the theory then unfortunately acquires some awkward features. Like fluid dynamics which began with linear theories, then proceeded to stable non-linear theories to describe turbulence, car-following theories have now been forced into some of the awkward mathematical complications which are associated with non-linear oscillations.

3.2 Mathematical Equations of Simple Follow-the-Leader Models

Follow-the-leader theory applies to single-lane dense traffic with no overtaking. For this theory to apply, the vehicles would normally be spaced at about 50 to 100 ft, or have a time-headway of about 6 to 8 sec. The theory is based on the assumption that each driver reacts in some specific fashion to a stimulus from the vehicle or vehicles ahead of and/or behind him. In general, road conditions do not affect interactions between vehicles, and so 'acceleration noise' is ignored in this theory. A different approach altogether is necessary for low concentrations where interactions between vehicles disappear.

Car-following theory attempts to describe mathematically the way vehicles go down a road, and to determine qualitatively what happens to the dynamics of this chain when there is a fluctuation in the motion, i.e. in the speed. Does the disturbance increase or die away? If the former is the case accidents may be caused, or a major bottleneck may develop, so this is important.

The basic equation of follow the-leader theory is not a dynamical, but a psychological one viz.

$$\text{Response} = \text{Sensitivity} \times \text{Stimulus}. \tag{3.1}$$

The response is taken as the acceleration of the following vehicle, since this is a quantity which is directly under the control of that driver. The stimulus could be a function of the position of a number of vehicles and their time-derivatives. It has been established, by experiment, that there is a high correlation between the response of a driver and the relative speed of his vehicle and the one ahead: thus the stimulus is taken as this relative speed.

Sensitivity has been taken by various authors, among them Herman and Potts (1961) and Edie (1961), as constant, inversely proportional to the spacing between consecutive vehicles or various other functions. A different approach was made by Newell (1962a); he took sensitivity and stimulus together. Of course all these expressions for the sensitivity have been made in an effort to account for experimental and observed data.

The stimulus-response equation of the follow-the-leader model is then:

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t)-\dot{x}_{n+1}(t)], \tag{3.2}$$

where x_n is the position of the nth vehicle;

T is the time lag of response to stimulus (this is of the order of 1·5 sec for 50% of drivers, and in the range 1·0 to 2·2 sec for all drivers);

λ is the sensitivity: this is also referred to by Herman *et al.* as the proportionality coefficient.

The following are possible functions which can be taken for the sensitivity:

(i) $\lambda = c$, constant.

(ii) $\lambda = \left.\begin{matrix} a, & s \leqslant s_c \\ b, & s > s_c \end{matrix}\right\}$

where $s = (x_n - x_{n+1})$ and s_c is a critical value of s. This is a step function.

(iii) $\lambda = c/s$, c constant. This is the case of reciprocal spacing.

(iv) $\lambda = \dot{x}_{n+1}/s^2$.

(v) $\lambda = c/s^2$.

All of the expressions (iii)–(v) can be included in the general expression.

$$\lambda = \frac{c\dot{x}^m_{n+1}(t+T)}{[x_n(t)-x_{n+1}(t)]^l}, \tag{3.3}$$

where l and m are constants. Choice of an expression of type (3.3) depends on how well it describes actual traffic phenomena, and two characteristics of the traffic flow are here of importance, stability and steady-state flow.

Consider first the question of stability. Stability criteria are difficult to establish except for the case $\lambda = c$. Then the chain is stable if, and only if, $\lambda T < 1/2$, and this can be regarded as providing a bound for the non-linear cases. It should be noted that the linearization of a non-linear system for small perturbations is possible only so long as the linearized system is stable. In this case any perturbation introduced in a stream of traffic remains small. Some numerical evaluation of certain simulated traffic problems indicate that this stability bound for the model given by equation (3.3) is very close to that of the linear model. A complete investigation of the stability of a non-linear system is not yet available although some work has already been done by Kometani and Sasaki (1958); Herman, Montroll, Potts and Rothery (1959); and Sasaki (1959).

Herman *et al.* investigated the question of stability by introducing a 'reaction-time', T, into a simple model. Putting $\dot{x}=u$, $\lambda=c$ in equation (3.2) and writing t for $(t+T)$ we get

$$\frac{\mathrm{d}u_{n+1}(t)}{\mathrm{d}t} = c[u_n(t-T)-u_{n+1}(t-T)]. \tag{3.4}$$

The complete solution of this equation can be accomplished by using the Laplace Transformation. A brief description of this method is given here: for further information see, for instance, Carslaw and Jaeger (1941).

Let $u_n^*(p)=\mathscr{L}[u_n(t)]=\int\limits_0^\infty u_n(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$ be the Laplace Transform of $u_n(t)$. A well-known result, derived by integration by parts states that

$$\mathscr{L}\left[\frac{\mathrm{d}u_n(t)}{\mathrm{d}t}\right] = pu_n^*(p)-u_n(0).$$

Also
$$\int\limits_0^\infty u_n(t-T)\,\mathrm{e}^{-pt}\,\mathrm{d}t = \mathrm{e}^{-pT}\int\limits_{-T}^\infty u_n(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$$
$$= \mathrm{e}^{-pT}\int\limits_0^\infty u_n(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$$

if $u(t)=0$ on the interval $(-T,0)$.

$$\text{i.e. } \mathscr{L}[u_n(t-T)]=\mathrm{e}^{-pT}\mathscr{L}[u_n(t)].$$

Using these results and the definition of a Transform in equation (3.4) we get.

$$[pu_{n+1}^*(p)-u_{n+1}(0)] = c\,\mathrm{e}^{-pT}[u_n^*(p)-u_{n+1}^*(p)]$$

i.e.
$$[p+c\,\mathrm{e}^{-pT}]\,u_{n+1}^*(p) = u_{n+1}(0)+c\,\mathrm{e}^{-pT}u_n^*(p).$$

The transformed solution, $u_{n+1}^*(p)$, is then given by

$$u_{n+1}^*(p) = \frac{u_{n+1}(0)+cu_n^*(p)\,\mathrm{e}^{-pT}}{p+c\,\mathrm{e}^{-pT}}. \tag{3.5}$$

The evaluation of $u_{n+1}(t)$ can be carried out by expanding the right-hand side and inverting term by term. For instance if $u(0)=1$ and $u(<0)=0$, the velocity of the following vehicle is given by

$$u_{n+1}(t) = 1-\sum_{j=0}^{N}\frac{[\alpha(j-N-\theta)]^j}{j!},$$

where $\alpha=cT$, $N=[t/T]$, the integral part of t/T and $\theta=t/T-N$.

The solution is, however, difficult to interpret in this form and an alternative approximate form is preferable. For this write equation (3.4) as

$$\frac{\mathrm{d}u_{n+1}(t)}{\mathrm{d}t} = c\frac{\mathrm{d}s_{n+1}(t-T)}{\mathrm{d}t}$$

and expand the right-hand side in a Taylor series as far as the term in T^2, thus

$$\frac{\mathrm{d}u_{n+1}(t)}{\mathrm{d}t} \approx c\left[\frac{\mathrm{d}}{\mathrm{d}t}s_{n+1}(t)-T\frac{\mathrm{d}^2}{\mathrm{d}t^2}s_{n+1}(t)+\frac{T^2}{2}\frac{\mathrm{d}^3}{\mathrm{d}t^3}s_{n+1}(t)\right].$$

Now since $$s_{n+1} = (x_n-x_{n+1})$$

then $$\frac{\mathrm{d}s_{n+1}}{\mathrm{d}t} = (u_n-u_{n+1}),$$

$$\frac{\mathrm{d}^2 s_{n+1}}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t}(u_n-u_{n+1}) \text{ etc.}$$

Hence, collecting terms in u_{n+1} on the left-hand side and those in u_n on the right-hand side, to order T^2,

$$\frac{cT^2}{2}\frac{\mathrm{d}^2 u_{n+1}}{\mathrm{d}t^2}+(1-cT)\frac{\mathrm{d}u_{n+1}}{\mathrm{d}t}+cu_{n+1} = \frac{cT^2}{2}\frac{\mathrm{d}^2 u_n}{\mathrm{d}t^2}-cT\frac{\mathrm{d}u_n}{\mathrm{d}t}+cu_n \quad (3.6)$$

This gives the velocity of the following vehicle, u_{n+1}, in terms of that of the one ahead, u_n, which is assumed known. Equation (3.6) is then a second order linear differential equation having a general solution made up of two parts, a complementary function and a particular integral. The complementary function is otherwise termed the transient. For this set the right-hand side of equation (3.6) equal to zero and assume as a solution $u_{n+1}(t)=\mathrm{e}^{mt}$ where m is to be determined. Differentiation and substitution in (3.6) then lead to the characteristic equation

$$m^2\left(\frac{cT^2}{2}\right)+m(1-cT)+c = 0$$

with roots
$$\frac{-(1-cT)}{cT^2} \pm \sqrt{\left(\frac{(1-cT)^2}{c^2T^4}-\frac{2}{T^2}\right)}.$$

Thus the transient will be non-oscillatory if

$$\frac{(1-cT)^2}{c^2T^4} > \frac{2}{T^2}$$

or
$$cT < (\sqrt{2}-1) = 0{\cdot}414.$$

If this is not satisfied the transient oscillates and stability depends on the sign of $(1-cT)$. The behaviour of the transient is therefore as follows:

$0 \leqslant cT < 0{\cdot}414$ non-oscillatory;

$0{\cdot}414 \leqslant cT < 1{\cdot}0$ damped oscillation;

$cT \geqslant 1{\cdot}0$ increasing oscillation.

These results are approximate and are different from those given by Herman *et al.*, which are exact. For the exact analysis consider equation (3.5). The general character of the solution is determined by the zeros of the denominator. These occur at the roots of

$$p+c\,\mathrm{e}^{-pT} = 0.$$

Putting $pT = \alpha + i\beta$ and equating real and imaginary parts, the roots can be found as the intersections of the curves

$$\begin{cases}\alpha+cT\,\mathrm{e}^{-\alpha}\cos\beta = 0,\\ \beta-cT\,\mathrm{e}^{-\alpha}\sin\beta = 0.\end{cases}$$

For large t the character of the solution depends on the pole having the largest real part, the contribution to the velocity from the other poles being heavily damped. Even for small t it can be shown that it is sufficient to consider this pole. The results are then

$cT \leqslant 1/\mathrm{e} = 0{\cdot}368$ non-oscillatory;

$1/\mathrm{e} < cT < \pi/2 = 1{\cdot}57$ damped oscillation;

$cT = \pi/2$ underdamped oscillation;

$cT > \pi/2$ increasing oscillation.

The type of instability described here, which is concerned with the response of one vehicle to its neighbour, has been called 'local instability'. It is also possible for a disturbance to grow as it is transmitted down a line of traffic, a phenomenon called 'asymptotic instability'. This is seen from the particular integral of (3.6) if it is assumed that $u_n = U_n \mathrm{e}^{i\omega t}$, for then it is found that, differentiating and substituting in (3.6),

$$U_{n+1}[-cT^2\omega^2/2+(1-cT)\,i\omega+c] = U_n[-cT^2\omega^2/2-cTi\omega+c]$$

and this implies that u_{n+1}/u_n grows as n increases if

$$\frac{1-cT}{cT} < 1 \quad \text{or} \quad cT > 1/2.$$

Note that if $cT \leqslant 1/\mathrm{e}$, the local stability criterion comes within the criterion for asymptotic stability. Hence the condition $cT \leqslant 1/\mathrm{e}$ implies the most desirable kind of response to a perturbation in the movement of the leading vehicle. For $1/\mathrm{e} < cT < 1/2$, the response is still highly damped although there is some oscillation. In practice, however, cT does not always satisfy the local stability criterion, and the result is to be seen in the kind of rear-end collisions encountered particularly on motor-ways.

Consider now the steady-state flow. For the equations of steady state (3·2) and (3·3) give for the stimulus-response equation

$$\ddot{x}_{n+1}(t+T) = \frac{c\dot{x}^m_{n+1}(t+T)[\dot{x}_n(t)-\dot{x}_{n+1}(t)]}{[x_n(t)-x_{n+1}(t)]^l}$$

or

$$\frac{\ddot{x}_{n+1}(t+T)}{\dot{x}^m_{n+1}(t+T)} = \frac{c[\dot{x}_n(t)-\dot{x}_{n+1}(t)]}{[x_n(t)-x_{n+1}(t)]^l}. \tag{3.7}$$

We can write equation (3.7) in its simplest form, putting $\dot{x}=u$, $l=m=0$ and omitting lag T, since we are considering the steady state, as

$$\frac{\mathrm{d}u_{n+1}}{\mathrm{d}t} = c(u_n-u_{n+1}).$$

Since

$$x_n - x_{n+1} = s_{n+1} = \frac{1}{k_{n+1}}$$

then

$$u_n - u_{n+1} = \frac{\mathrm{d}s_{n+1}}{\mathrm{d}t} = \frac{-1}{k^2_{n+1}}\frac{\mathrm{d}k_{n+1}}{\mathrm{d}t}, \tag{3.8}$$

where s_{n+1} is the spacing and k_{n+1} the concentration. The equation can then be written, omitting the subscript, as

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{-c}{k^2}\frac{\mathrm{d}k}{\mathrm{d}t}.$$

Integrating once gives

$$u = \frac{c}{k}+A.$$

This equation says nothing useful about the steady state, merely that then the relative velocity is zero. If it is assumed that all the changes of concentration are governed by this equation, and that when the concentration is k_j (jam concentration, i.e. about 350 vehicles/mile if an average vehicle length is taken to be 15 ft) the flow, and therefore the velocity is zero, we can evaluate A by using the boundary condition $u=0$ at $k=k_j$ to give $A=-c/k_j$ and then

$$u = c(1/k-1/k_j) = c(s-s_j).$$

Now in the steady state $q = uk,$

and hence

$$q = c(1-k/k_j) \tag{3.9}$$

gives the flow-concentration relationship. Note that c is here a flow rate. The quantity k/k_j is the normalized concentration, and is sometimes written η.

Similarly, putting $\dot{x}=u$ and $l=1$, $m=0$ i.e. $\lambda=c/(x_n-x_{n+1})$, and omitting lag T, equation (3.7) becomes

$$\frac{\mathrm{d}u_{n+1}}{\mathrm{d}t} = \frac{c(u_n-u_{n+1})}{(x_n-x_{n+1})}. \tag{3.10}$$

Equation (3.10) can be written, by virtue of (3.8) as

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{-c}{k}\frac{\mathrm{d}k}{\mathrm{d}t} \tag{3.11}$$

which integrates to give

$$\begin{aligned} u &= c\ln\frac{k_j}{k} \\ &= c\ln\frac{s}{s_j}. \end{aligned}$$

In the steady state $q = uk$ and so the flow-concentration relationship becomes

$$q = ck \ln \frac{k_j}{k}, \tag{3.12}$$

where c is here a velocity.

Equations (3.9) and (3.12) are not valid near $k = 0$, so that the constant of integration must be found, in both cases, by using the boundary condition given above, and the equations apply to dense traffic only. Herman *et al.* have shown experimentally that the assumptions leading to equation (3.10) are reasonable and that equation (3.12) may be fitted to their steady-state data.

Putting $l = 2$, $m = 1$, i.e. $\lambda = c\dot{x}_{n+1}/(x_n - x_{n+1})^2$ in equation (3.7) gives

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -cu \frac{\mathrm{d}k}{\mathrm{d}t}$$

and on integrating,

$$\ln u = -ck + A.$$

This last equation, unlike the two previous ones, is not valid near $k = k_j$ (in other words this case refers to light traffic), so A must be evaluated using the other boundary condition $u = u_f$ at $k = 0$. Hence

$$\begin{aligned} u &= u_f \mathrm{e}^{-ck} \\ &= u_f \mathrm{e}^{-c/s} \end{aligned}$$

and then

$$q = u_f k \, \mathrm{e}^{-ck}, \tag{3.13}$$

where c is here a spacing.

Finally putting $l = 2$, $m = 0$, i.e. $\lambda = c/(x_n - x_{n+1})^2$ in equation (3.7) gives

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -c \frac{\mathrm{d}k}{\mathrm{d}t}$$

and on integration

$$u = -ck + A.$$

But again $u = u_f$ at $k = 0$ so that

$$u - u_f = -ck = -c/s$$

and

$$q = k(u_f - ck). \tag{3.14}$$

It will be noticed that equations (3.9) and (3.12) are arrived at after making the assumptions that there is a single relationship between q and k for all possible values of k, and that q is zero when $k = k_j$.

The first of these is that in the steady state

$$q = uk = f(k)$$

or

$$u = \phi(s). \tag{3.15}$$

If it is assumed that a similar equation holds for unsteady flow, then (3.15) gives, on differentiation,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \phi'(s)\frac{\mathrm{d}s}{\mathrm{d}t}$$

which is equivalent to (3.11). In fact (3.11) is really

$$\frac{\mathrm{d}u}{\mathrm{d}k} = -\frac{c}{k},$$

and this is merely to assume an expression for $\mathrm{d}u/\mathrm{d}k$, which is not very different from assuming a relationship between u and k directly. Equation (3.11) shows that if there is a steady-state relationship the acceleration of a vehicle must depend on the relative velocity compared with the vehicle ahead and that this dependence decreases as s increases.

Returning now to equation (3.7), in general, integration yields a steady-state flow equation of the form

$$f_m(u) = cf_l(s) + A, \tag{3.16}$$

where u is the steady state speed of a stream of traffic,
s is the constant average spacing,
c, A are some constants consistent with physical restrictions,
and the function f_p for $p = m, l$ is given by

$$f_p(x) = \begin{matrix} x^{1-p}, & p \neq 1 \\ \ln x, & p = 1. \end{matrix}$$

The integration removes the time-lag T in so far as the steady-state equations are concerned. The constant of integration A is obtained from a boundary condition viz. either $u = u_f$ (free speed) at $k = 0$, or $u = 0$ at $s = s_j$ (jam spacing), the boundary condition to be used depending on the values of m and l under consideration.

Using equation (3.16) and the definition of flow $q = uk$, $k = 1/s$ or equations (3.12) and (3.14) we can obtain relationships between the speed u, concentration k and flow q, and thus plot flow-concentration curves. For convenience the relevant flow-concentration equations are summarized below.

(a) $m=0$, $l=1$, i.e. $\lambda=c/s$, $q=ck\ln(k_j/k)$.
Maximum given by $\mathrm{d}q/\mathrm{d}k=0$, i.e. $k_j/k=\mathrm{e}$, and then $q=ck_j/\mathrm{e}$. Hence, if q/ck_j is plotted against k/k_j the maximum of the curve is at $(1/\mathrm{e}, 1/\mathrm{e})$.

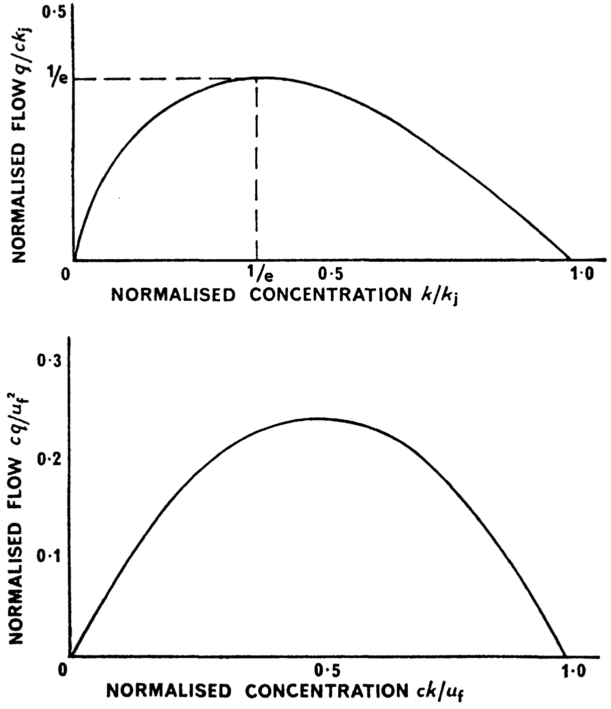

FIG. 3.1. Flow-concentration curves $f_m(u)=f_l(s)+A$ for $m=0$ and $l=1, 2$.

(b) $m=0$, $l=2$, i.e. $\lambda=c/s^2$, $q=k(u_f-ck)$.
This is a parabola with a maximum at $k=u_f/2c$, and then $q/u_f=u_f/4c$. Hence if cq/u_f^2 is plotted against ck/u_f the maximum of the curve is at $(1/2, 1/4)$.

Curves corresponding to (a) and (b) are plotted in Figure 3.1.

Referring now to equation (3.2) viz.

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t)-\dot{x}_{n+1}(t)],$$

it should be noted that if traffic is moving at a constant speed this equation is satisfied whatever the spacing, hence no unique equation of steady state is implied by it. However, a transition from one steady state to another, after a perturbation in the leading car, can be obtained by a process which is governed by equation (3.2): this process is the integration of (3.2) on both sides, a procedure which eliminates the lag T and yields a relationship between the speed and the spacing of vehicles in a steady-state stream. This transition must lead to the right boundary conditions, i.e. either zero flow at jam concentration or free speed at zero concentration. However, any steady-state condition is actually obtained through a transition of this kind, and it may in practice include the boundary condition. It seems reasonable therefore, to accept the steady-state flow relationship as being implied by the basic stimulus-response equation (3.2). But it should be noted that it is possible for steady-state streams to exist which are not governed by this law.

3.3 Alternative Derivations of the Steady-state Equations

The steady-state equations can be derived from other considerations. Suppose that while driving at a speed u every driver tries to maintain a distance between his car and the next one ahead equal to, say,

(his own stopping distance) –
(an assumed stopping distance of the leading car),

where the stopping distances are defined by, for instance, the Highway Code.

Then the spacing, s, is given by the equation

$$s = s_0+Au+Bu^2, \tag{3.17}$$

where s_0, A, B are certain constants, e.g. s_0 may be the length of each headway measured from front bumper to front bumper, A a time-lag reaction to a sudden braking of the leading car and $B=(1/2d_{n+1}-1/2d_n)$, where d_n is the desired maximum deceleration of the nth or leading car, which probably has to be assumed the maximum physically possible. Kometani and Sasaki (1961) have investigated such a quadratic relation between spacing and speed.

Substitution for u, obtained from (3.17), into $q=uk$, $k=1/s$ yields a relationship between flow and concentration thus

$$u = \frac{-A \pm \sqrt{[A^2-4B(s_0-s)]}}{2B}$$

$$= \frac{\alpha}{2\beta}\left\{-1+\left[1+\frac{4\beta}{\alpha^2}\left(\frac{k_j}{k}-1\right)\right]^{1/2}\right\},$$

where $k_j=1/s_0$, $\alpha=A/s_0$, $\beta=B/s_0$. Hence

$$q = q_0\left\{\frac{-k}{k_j}+\left[\frac{k^2}{k_j^2}+\gamma\frac{k}{k_j}\left(1-\frac{k}{k_j}\right)\right]^{1/2}\right\}$$

where $$q_0 = \alpha k_j/2\beta, \quad \gamma = 4\beta/\alpha^2.$$

Maximum flow is obtained for $\mathrm{d}q/\mathrm{d}k=0$, i.e. for

$$\frac{1}{k_j} = \frac{\frac{1}{2}\left[\frac{2k}{k_j^2}+\frac{\gamma}{k_j}-\frac{2\gamma k}{k_j^2}\right]}{\left[\frac{k^2}{k_j^2}+\frac{k}{k_j}\left(1-\frac{k}{k_j}\right)\right]^{1/2}}.$$

Squaring, cross-multiplying and simplifying gives

$$\left(\frac{k}{k_j}\right)^2\gamma(1-\gamma)-\left(\frac{k}{k_j}\right)\gamma^2+\frac{\gamma^2}{4} = 0$$

and hence $$\frac{k}{k_j} = \frac{(\sqrt{\gamma}-\gamma)}{2(\gamma-1)},$$

neglecting the negative root.

Then $$q_{\max} = \frac{q_0\gamma}{2(1+\sqrt{\gamma})} = \frac{k_j}{(\alpha+2\sqrt{\beta})}.$$

Writing $Q=q/q_{\max}$ and $K=k/k_j$ we get

$$Q = \frac{2(1+\sqrt{\gamma})}{\gamma}\{-K+[K^2+\gamma K(1-K)]^{1/2}\}$$

and so $$\frac{\mathrm{d}Q}{\mathrm{d}K} = \frac{2(1+\sqrt{\gamma})}{\gamma}\left\{-1+\frac{1/2[2K+\gamma-2\gamma K]}{[K^2+\gamma K(1-K)]^{1/2}}\right\}.$$

As K (or k)→0, $\mathrm{d}Q/\mathrm{d}K$ (or $\mathrm{d}q/\mathrm{d}k$)→∞. This is all right because interaction between vehicles is statistically weak at low concentrations, and as was mentioned earlier, follow-the-leader techniques do not apply. If Q is plotted against K it can be seen that near zero concentration the flow-concentration curve is a straight line making an angle of just less than 90° with the vertical.

Differentiation of (3.17) gives

$$\dot{s} = (A+2Bu)\dot{u}$$

or

$$\dot{u} = \frac{c_1\dot{s}}{1+c_2 u},$$

which could be regarded as a stimulus-response equation, possibly after the introduction of a time-lag. This gives a third way of arriving at both steady-state and stimulus-response laws, in this case from elementary considerations.

Experimental data can, and indeed must, be used to determine which one, if any, of the above models agrees best with the real situation. Correlation studies have been used in an attempt to determine m and l, or at least to determine whether they are positive or negative; also to obtain the relative magnitudes of s_0, A, and B in equation (3.17). Not much has come out of all this except that probably $0<l<2$. It is now fairly well agreed, however, that the best functions for the sensitivity are $\lambda = u/s^2$, $1/s$ or $1/s^2$ with the balance in favour of $1/s$. It must be noted also that no strong arguments can be given in favour of any single equation of steady state.

Further discussion of flow-concentration relationships is deferred until a later chapter after kinematic theories have been considered.

3.4 Extensions of the Stimulus-Response Equation

There are two possible ways in which the stimulus-response equation (3.2) can be extended. It is possible that a following vehicle responds to vehicles other than the one immediately in front – the vehicle next but one ahead for instance. In this case the equation would be

$$\ddot{x}_{n+2}(t+T) = \lambda_1[\dot{x}_{n+1}(t)-\dot{x}_{n+2}(t)]+\lambda_2[\dot{x}_n(t)-\dot{x}_{n+2}(t)]. \quad (3.18)$$

It can be shown that the effect of the addition of the extra term is, to the first order, equivalent to replacing λT by $[\lambda_1 T+\lambda_2 T]$, so that the sensitivity of control is increased. Experiment and analysis in the United States both suggest that the driver follows only the vehicle immediately ahead, or at least that this is the most significant aspect of car-following. Results are not absolutely conclusive, but it is probably not worthwhile to consider these more complicated models.

The suggestion that decelerations and accelerations might not follow the same law has been made by several authors, for example Newell (1962), who noticed a slower response of drivers to accelerations than to decelerations. Herman and Potts reported that measurements of their sensitivity constant, which is essentially the slope of the velocity-headway curve, gave slightly different values for accelerations and decelerations; but they did not think the difference was large enough to be significant. It was not realized at the time that even a small difference could lead to irreversibility or instability. More recently an effort has been made to separate velocity-headway data according to whether the vehicles are accelerating or decelerating, and significantly larger spatial headways have been found during acceleration. It should be noted also that vehicles have different capacities for acceleration and deceleration, and acceleration waves propagate more slowly than deceleration waves. Most of the experimental evidence is, however, indirect.

We shall give here only a general description of the implications of this theory, which is developed in detail in a paper by Newell (1963). On this theory, not one λ, but two asymmetrical λ's viz. λ_+, λ_-, exist, one where the relative speeds between adjacent vehicles are positive, and one where they are negative, and such that $\lambda_- > \lambda_+$.

The conclusions to be drawn from Newell's theory for a deceleration are the same as those of any car-following or continuum theory, i.e. that the velocity-headway curve is concave, and by virtue of this the deceleration wave develops into a shock. This means that the wave eventually acquires a stable form which propagates through the line of vehicles with no change in structure. In the simpler theories of §3.2 the velocity-headway curve is concave for accelerations also, and an expanding wave is always produced. This means that the wave corresponding to the lowest vehicle-velocity propagates with a slower forward wave velocity (or a more negative velocity if it propagates backwards) than the waves of higher vehicle-velocity. The further back one goes from the leading vehicle the slower is the rate of acceleration, in contrast with a shock, for which the rate of acceleration or deceleration approaches some constant value.

The main qualitative difference between the theory of §3.4 and that of §3.2 arises because, in fact, the shape of this curve is not concave everywhere. Convex portions of the curve correspond to acceleration shocks in the same way that concave portions correspond to deceleration shocks.

Instability arises in this theory because acceleration waves do not necessarily propagate with the same velocity. The result is that drivers can co-operatively induce each other to enlarge or reduce the headways between vehicles.

In a stable continuum theory fluctuations are always dissipated in the following way. If a vehicle accelerates and then decelerates back to its original velocity it sends out two groups of waves. The acceleration waves expand and the deceleration waves form a shock: but the acceleration waves have a range of velocities which include the shock velocity. This implies that the shock always overtakes at least part of the expanding wave and the two groups of waves proceed to annihilate each other systematically.

In the later theory waves for two opposing effects do not necessarily travel with the same velocity. If, in particular, the acceleration wave which is created first propagates more slowly than the deceleration wave which is supposed to compensate for it, the compensating wave will never catch up and the disturbance will not be dissipated. There is a drift to higher velocities, which is finally halted by a deceleration shock.

A perturbation from the same initial state consisting of a deceleration followed by an acceleration, on the other hand, does not cause an analogous effect becsuse the two waves of deceleration and acceleration travel with essentially the same velocity and tend to annihilate each other. There is a drift to lower velocities: in this case the drift is not halted by a shock, in fact the drift is irreversible and persists until the velocity approaches zero.

The more complicated theory has explained the flow in tunnels whereas the former did not. It has explained the instability and the propagation of stopping waves behind a bottleneck, together with some of the secondary effects associated with it. Altogether about a dozen experimentally observed effects which are qualitatively inconsistent with simpler theories of stable traffic flow have a possible interpretation in terms of this later theory. The theory of §3.2, for instance, predicts that behind a bottleneck an increase in capacity should normally be accompanied by a decrease in velocity whereas experiments have shown the opposite.

Both the theory and the experiments are still, however, in a rather primitive state. There is a shortage of even qualitative descriptions of the evolution of traffic patterns over large distances in contrast with data taken at isolated points. The theory is still also incomplete: although there is a mechanism for instability in the theory, there is no explicit method for evaluating the drift in velocities, which has been supposed to arise from random fluctuations about which nothing is yet known.

We noted in §3.2 that in follow-the-leader theory acceleration noise was ignored. In general, an acceleration noise component $\beta(t)$ should be added. The random function $\beta(t)$ is determined by a distribution function $f(x)$ which is such that $f(x)$ is the probability that $\beta(t)$ has a value in the range $(x, x+\delta x)$ at time t. In stable smooth-flowing traffic the effect of $\beta(t)$ dies out as it is propagated down the line of vehicles, and usually disappears completely by the time it reaches the fifth or sixth from the one considered.

The theory described in this section has contributed to a general understanding of the situation but we are still a long way from having a theory that can be put to practical use by traffic engineers. It is obvious that there are considerable difficulties involved in setting up a completely realistic car-following theory. The situation is extremely complex, and the question arises as to whether it might be better to overlook microscopic details of the motion of individual vehicles and instead look at the problem macroscopically, i.e. treat the flow as a continuum. The flow-concentration curve is then to be regarded as an experimental result on which to build a theory of flow.

The theory discussed in this chapter proceeds from detailed assumptions concerning the behaviour of drivers, whereas that of Lighthill and Whitham (1955), whose theory is discussed in the next chapter, is an endeavour to derive results from the minimum of hypotheses. It is not able to yield precise results in all fields – it will not handle the matter of queue-length distributions, for example – but it complements car-following theory by providing broad overall results.

CHAPTER 4

Kinematic Theories and Fluid Analogies

4.1 Introduction

This chapter contains an outline of a theory of traffic flow based on a quite different idea from that of car-following, namely that of the analogy between the flow of traffic and that of a fluid. The theory was originally suggested by flood movements in rivers and by the flow about supersonic projectiles. Before describing the theory it must be admitted that there are a number of objections to this sort of analogy. In the first place, in a fluid action and reaction are opposed whereas in traffic impulses are generally only transmitted in one direction. Nor are the laws of mass, force and compressibility applicable. In the second place the acceleration and deceleration of vehicles have finite limits. In the third place the behaviour of traffic arises from its statistical properties. More will be said in later chapters about this. The continuous flow approach represents the limiting behaviour of a stochastic process for a large population (the total number of vehicles) and is applicable in the main, therefore, to large scale problems only i.e. to the distribution of traffic on long crowded roads. It has some relevance, however, as will be explained later, to the starting flow at junctions.

4.2 The Continuum Theory

The continuum theory of traffic flow is based on two relations. One is the equation of continuity, which can be otherwise expressed as the law of conservation of vehicles. The other, which is quite fundamental, is an assumed relation between the flow q, which measures quantity per unit time, and the concentration k, which measures quantity per unit distance. In one-dimensional flow, if we assume such an approximate functional relation at each point between these two quantities, and in general the position x, the wave property follows directly from the equation of continuity satisfied by q and k. These waves were called 'kinematic' by Lighthill and Whitham (1955) in order to distinguish them from the wave motions encountered in dynamical systems from which they are physically quite distinct. These classical 'dynamic' waves depend not only on the equation of continuity but also on Newton's second law of motion together with some assumptions relating a stress to a displacement, strain or curvature.

If we take measurements of q at two point a short distance Δx apart the excess of q on the down-stream side must equal the diminution of vehicles in the stretch of road between the two points. Hence the law of conservation of vehicles gives

$$\frac{\partial q}{\partial x}\Delta x = -\frac{\partial}{\partial t}(k\Delta x)$$

so that

$$\frac{\partial k}{\partial t}+\frac{\partial q}{\partial x} = 0. \tag{4.1}$$

This equation is of the first order only, so that kinematic waves possess only one wave velocity at each point whilst dynamic waves possess at least two, one forward and one backward relative to the medium.

In general both q and k are functions of both distance and time so that $q=f(x,t)$, $k=\phi(x,t)$ and $q=q(k,x)$. This latter relation is certainly approximate. For one thing there is always the intervention of a time-lag between adjustments of flow and concentration at a given point. For another it may be valid only in the statistical sense. This treatment is accurate, therefore, only over times which are large compared with the time-lag, and only if deviations from the mean q-k relation are small compared with the wave effects. However, first assume that q is a function of concentration only so that $q=q(k)$. In this case

$$\frac{\partial q}{\partial x} = \frac{\mathrm{d}q}{\mathrm{d}k}\cdot\frac{\partial k}{\partial x}.$$

Substitution in (4.1) gives

$$\frac{\partial k}{\partial t}+\frac{\mathrm{d}q}{\mathrm{d}k}\cdot\frac{\partial k}{\partial x} = 0. \tag{4.2}$$

For a given function $q(k)$, this is a partial differential equation giving the variations of k (and hence q) with time and space, and can be regarded as an initial value problem.

In order to understand equation (4.2) first consider small variations of k about an average value. In this case we can take

$$\frac{\mathrm{d}q}{\mathrm{d}k} = c, \text{ constant}$$

so that (4.2) becomes

$$\frac{\partial k}{\partial t} + c\frac{\partial k}{\partial x} = 0.$$

This equation can be solved in the usual way. It has one system of characteristics given by

$$\mathrm{d}x = c\,\mathrm{d}t$$

and the solution is $$k = \phi(x-ct),$$

i.e. k (and hence q) is a function of $(x-ct)$. In other words the variations of k and q travel with speed c. The flow q remains constant on each kinematic wave, k and c are constant along it and the wave moves with constant velocity. The flow-concentration curves in this case are straight lines. The mean velocity of individual vehicles at a point is $u = q/k$ and this differs in concept and value from the wave velocity c.

Now since $u = \dfrac{q}{k}$,

$$c = \frac{\mathrm{d}q}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}(uk) = u + k\frac{\mathrm{d}u}{\mathrm{d}k}.$$

The case $c > u$ corresponds to a mean velocity which increases with concentration, the case $c < u$ corresponds to a mean velocity which decreases with concentration. The former is the case applicable to flow in rivers, the latter to traffic flow. The speed of individual vehicles is, of course, necessarily non-negative; the wave velocity c, however, will be negative if the road is used beyond capacity, and hence in dense traffic, conditions propagate backwards along the road.

For the case of non-uniform flow, $q = q(k, x)$ and

$$c = \left(\frac{\partial q}{\partial k}\right)_{x\,\text{constant}}$$

$$= c(k, x) \text{ or } c(q, x). \qquad (4.3)$$

In this case the paths of the waves are not straight lines; the curves are in fact given by

$$t = \int_0^x \frac{\mathrm{d}x}{c(q,x)} + \text{constant}$$

$$= \int_0^x \left(\frac{\partial k}{\partial q}\right) \mathrm{d}x + \text{constant}.$$

In order to evaluate this expression the integral

$$\int_0^x k(q,x)\,\mathrm{d}x$$

is calculated for different values of q and then differentiated with respect to q under the integral sign. The wave velocity is then the slope of the flow-concentration curve for fixed x.

For a full analysis of equation (4.2) the independent variables are changed from x and t to k and T, where $T = t$. Since k is a function of x and t

$$\mathrm{d}k = \frac{\partial k}{\partial x}\mathrm{d}x + \frac{\partial k}{\partial t}\mathrm{d}t$$

$$= \frac{\partial k}{\partial x}\left(\frac{\partial x}{\partial k}\mathrm{d}k + \frac{\partial x}{\partial T}\mathrm{d}T\right) + \frac{\partial k}{\partial t}\left(\frac{\partial t}{\partial k}\mathrm{d}k + \frac{\partial t}{\partial T}\mathrm{d}T.\right)$$

Hence, putting $T = t$

$$\mathrm{d}k = \left(\frac{\partial k}{\partial x}\frac{\partial x}{\partial k} + \frac{\partial k}{\partial t}\frac{\partial t}{\partial k}\right)\mathrm{d}k + \left(\frac{\partial k}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial k}{\partial t}\right)\mathrm{d}t.$$

Comparing coefficients

$$1 = \frac{\partial k}{\partial x}\frac{\partial x}{\partial k} + \frac{\partial k}{\partial t}\frac{\partial t}{\partial k},$$

$$0 = \frac{\partial k}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial k}{\partial t}.$$

Solving these equations for $\partial k/\partial x$ and $\partial k/\partial t$,

$$\frac{\partial k}{\partial x} = \frac{1}{J},$$

$$\frac{\partial k}{\partial t} = \frac{-1}{J}\frac{\partial x}{\partial t},$$

where J is the Jacobian $\partial(x,t)/\partial(k,t)$. Substituting in equation (4.2)

$$\frac{\partial x}{\partial t} - q'(k) = 0$$

so that

$$x = \phi(k) + tq'(k),$$

$\phi(k)$ being an arbitrary function of k representing the initial conditions at $t=0$. This equation implies that each value of k propagates with speed $q'(k)$ as in the simpler analysis above.

A number of different cases can be distinguished according to the behaviour of k. Suppose first that k decreases gradually in the forward direction. Since the lower values of k travel faster than the higher values of k, this means that the front region travels faster than the rear region, and the decrease in k is gradually spread out. If all the values of k correspond to less than the capacity of the road, everything propagates forwards; otherwise at least some of the conditions travel backwards.

Secondly, suppose there is a sharp decrease in k, as in the case of a traffic light. If the light is at red for long enough, then the road will be packed behind the traffic light and empty for a long way in front of it. When the light changes to green (i.e. at time $t=0$ when the equation begins to apply) there is an abrupt transition of k from its jam value k_j behind the light to its zero value in front of the light. Corresponding to this abrupt change, we could say that initially all values of k coexist at the traffic light. Afterwards each value of k propagates with its own characteristic velocity. Values of k corresponding to higher than capacity flow travel backwards; values of k corresponding to less than capacity flow travel forwards; and the value of k corresponding exactly to capacity flow does not travel at all, since for this value, $q_{\max}$, $q'(k)=0$. The fastest forward speed corresponds to the tangent at $k=0$, i.e. to the speed of vehicles seeing a clear road in front of them. The fastest backward speed corresponds to the tangent at the maximum value of k, and this is the boundary region rearwards along which the jam resolves, i.e. along which vehicles begin to move, since this alone is stationary. At the traffic light itself the value of k will be the one corresponding to that of capacity flow, the one for which $q'(k)=0$. Accordingly, the capacity of the road can readily be measured by installing a traffic light and allowing jam conditions to build up behind it before turning it to green. If R is the length of the red phase, G that of the green, and if, on average, vehicles arrive at the junction at a rate q and leave at a rate q_1, the total number of vehicles arriving in a complete cycle is

$$q(R+G).$$

They will all pass across the junction during the succeeding green only if

$$q(R+G) < q_1 G.$$

Hence the capacity of the junction is

$$\frac{G}{R+G} q_1.$$

This method of measuring the capacity of a junction has been one of the most valuable results of this theory to the practising traffic engineer, and shows, incidentally, what an efficient means of traffic regulation a traffic light is.

Thirdly let us suppose that k increases in the forward direction. This initial situation obviously leads to mathematical difficulties after a short time, because the rearward conditions corresponding to low concentration travel faster and collide with the front conditions corresponding to high concentration. The only thing that can happen now is that a shock-wave builds up, i.e. there is a sudden transition from light traffic conditions in the rear to heavy traffic condition in front. The speed U of such a discontinuity, or shock-wave, can readily be evaluated by using the same conservation considerations as were used in deducing continuous kinematic waves. If k_1, q_1 describe the conditions in front of the shock-wave, and k_2, q_2 behind it, then the rate at which vehicles emerge from the shock-wave from the front is $q_1 - k_1 U$, whereas the rate at which they enter the shock-wave from the rear is $q_2 - k_2 U$. By the law of conservation of vehicles these two must be equal, and therefore the speed of the shock-wave is given by

$$U = \frac{q_2 - q_1}{k_2 - k_1}. \tag{4.4}$$

This is clearly the inclination of the chord joining the two points on the flow-concentration curve for a given x which correspond to the states ahead of and behind the shock-wave when it reaches x. In the limit, when the shock becomes a continuous wave, the slope of the chord becomes the slope of the tangent, and the velocity given by (4.4) coincides with that given by (4.3).

The theory shows, therefore, that the front of a heavy concentration of traffic tends to smooth out, whereas the rear sharpens up and forms a jam. This means that a vehicle enters heavy traffic conditions suddenly but runs out of them gradually. This fits in perfectly with experience.

The theory can easily be generalized to include bottlenecks. A bottleneck is a place where a road of a certain capacity has a section on it of a lower capacity, e.g. a village situated on a trunk road. If a bulge of traffic moves into the bottleneck, jam conditions will spread backwards as a shock-wave along the road, and all such examples can be worked through graphically.

The simplicity of kinematic waves is essentially due to the fact that they travel in one direction only, whereas the usual second-order wave equation describes waves travelling in both directions. On the other hand, kinematic waves being, in general, non-linear they show such interesting phenomena as group velocity and the formation of shock-waves. These occur in general for non-linear wave problems, but cannot usually be so easily isolated as in this case.

4.3 More Complicated Models

Equation (4.2) gives no indication of possible instability: this is not surprising since this equation contains no time-dependence. However it has been noted above that the behaviour of individual vehicles is likely to depend on both the space and time derivatives of k. Thus it is probably more realistic to write

$$q = q\left(k, \frac{\partial k}{\partial x}, \frac{\partial k}{\partial t}\right).$$

Thus

$$\begin{aligned}\frac{\partial q}{\partial x} &= \frac{\partial q}{\partial k}\frac{\partial k}{\partial x}+\frac{\partial q}{\partial k_x}\frac{\partial k_x}{\partial x}+\frac{\partial q}{\partial k_t}\frac{\partial k_t}{\partial x}\\ &= c\frac{\partial k}{\partial x}+\mu\frac{\partial^2 k}{\partial x^2}+\nu\frac{\partial^2 k}{\partial x\,\partial t}\end{aligned}$$

where $$k_x = \frac{\partial k}{\partial x}, \quad k_t = \frac{\partial k}{\partial t}.$$

Equation (4.1) now becomes

$$\frac{\partial k}{\partial t}+c\frac{\partial k}{\partial x}+\nu\frac{\partial^2 k}{\partial x\,\partial t}+\mu\frac{\partial^2 k}{\partial x^2} = 0, \tag{4.5}$$

a form similar to that suggested, but not investigated, by Lighthill and Whitham. The extra terms refer to two higher order effects, a 'diffusion' effect and an 'inertia' effect. Consider first the diffusion effect. Since each driver is concentrating primarily on the road immediately ahead of him he adjusts his speed in accordance with his local concentration. This means that the flow q is dependent not only on the concentration but also on the concentration gradient. This leads to a 'diffusion' which spreads out the shock-waves: drivers do not in fact reduce speed instantaneously at a shock because they see it coming. The inertia effect is due to the fact that a driver must apply either the accelerator or the brake in order to reach the desired speed, and this means a time-lag. When these higher order effects are added they lead to equation (4.5) in which ν refers to an initial time constant for adjustments of speed and μ refers to a diffusion coefficient or decrement of flow for unit concentration gradient. This equation is similar to the equation governing waves in rivers when higher order effects are taken into account.

The propagation of waves may be examined by assuming that k varies like $\exp(\gamma t+i\beta x)$ where β is real, and if such solutions exist then, substituting in (4.5),

$$\gamma+ic\beta+\nu\gamma i\beta-\mu\beta^2 = 0$$

i.e.

$$\gamma = \frac{(\mu\beta^2-ic\beta)(1-i\nu\beta)}{1+\beta^2\nu^2}$$

$$= \frac{\beta^2(\mu-\nu c)-i(c-\mu\nu\beta^2)}{1+\beta^2\nu^2}$$

From this it is seen that if the amplitude of a wave is not to grow with time then

$$c > \frac{\mu}{\nu}.$$

If μ and ν are of the same sign, this implies that c cannot become negative. There is some evidence to support this idea for it has been noted that in tunnels, as the flow approaches its maximum value (at which c is zero), the traffic stream does tend to oscillate. As a result of these oscillations, gaps appear in the traffic and immediately after a gap the flow is found to be higher than normal. Further, after the gap has passed and the concentration has built up, the flow falls off. Behind a gap the value of $\partial k/\partial x$ will be negative; as the concentration builds up $\partial k/\partial t$ is positive. Thus these observations seem to indicate that $\partial q/\partial k_x$ is negative, since the flow is greater when $\partial k/\partial x$ is negative, and also that $\partial q/\partial k_t$ is negative, since the flow decreases when $\partial k/\partial t$ is positive, which is what was suggested above. This phenomenon corresponds to that of the so-called 'roll-waves' in rivers in which uniform flow is unstable and tends to degenerate into a succession of rapid accelerations and decelerations. In traffic this can be seen sometimes in convoys of vehicles. The behaviour of flow behind a controlled junction could, in principle, be evaluated on the basis of equation (4.5).

The reason for the instability can be seen if we make a first approximation

$$\frac{\partial k}{\partial t} \approx -c\frac{\partial k}{\partial x}$$

and substitute in equation (4.5) to give

$$-\frac{\nu}{c}\cdot\frac{\partial^2 k}{\partial t^2}+\mu\frac{\partial^2 k}{\partial x^2} = 0$$

or

$$\frac{\partial^2 k}{\partial t^2} = \frac{\mu c}{\nu}\cdot\frac{\partial^2 k}{\partial x^2}. \tag{4.6}$$

This has the same form as the usual second-order wave equation. The effect of μ and ν is to produce waves travelling with velocity $\pm\sqrt{(\mu c/\nu)}$ and the condition $c>\mu/\nu$ merely means that the kinematic wave velocity c must be greater than the 'elastic' wave velocity $\sqrt{(\mu c/\nu)}$.

4.4 Connection with Follow-the Leader Theories

Two papers by Franklin (1962) and one by Newell (1962) link car-following with the Lighthill–Whitham theory by showing that the existence of waves in traffic flow can be deduced from car-following theories as well as from the continuous flow analogy. In the last analysis all the different theories of traffic must in some way be unified. As yet only small beginnings have been made in this direction and efforts are still in the main piecemeal. The problem of passing from the simple but effective hydrodynamic analogy to a molecular (stochastic) theory is as difficult in the case of traffic as in the similar case of a liquid. The problem has not yet been solved in either context. This matter of unification will be pursued in the next chapter.

4.5 Boltzmann-like Theory of Traffic Flow

A paper by Prigogine, Herman and Anderson (1963) suggests the following approach, which is based upon the fact that certain analogies can be made between the flow of vehicles on a road and the flow of molecules in a gas as described in the kinetic theory of gases. In Boltzmann's classical theory it is assumed that molecules move about freely in space most of the time, but that occasionally their motion is disturbed by collisions with other molecules. On a road having a low traffic volume a vehicle moves freely most of the time but occasionally there is some kind of interaction with another vehicle, the result of which is usually that one vehicle overtakes another. This is the common feature which suggests that an analogy between the two situations might bear fruitful results.

There are both significant differences and similarities between vehicles overtaking on a road and molecules colliding. In the first place the gas molecules are free to move in three dimensional space whereas vehicles move (essentially) in only one dimension. It is also true that the laws governing collision in a gas are essentially those of particle dynamics and are already well understood: they are fundamentally deterministic, since if the exact positions and velocities of the molecules are known, the entire collision process is specified. The situation of interacting vehicles, on the other hand, is not, as yet, well understood, and can really only be described in a statistical way.

However there is one similarity: if an overtaking vehicle is thought of as imparting its own excess speed to the vehicle in front whilst assuming the slower speed itself, we can regard this as an elastic collision in which the two vehciles have changed identities. In this way overtaking can be regarded as a 'collision' in which both momentum and energy are conserved, at least for the initial and final states. There are some difficulties here. In the case of particle dynamics momentum and energy are conserved at all times, and this means that the average velocity of each particle is constant and independent of the number of collisions. This is obviously not true for traffic.

In spite of these differences there seem to be some grounds for optimism, for the following reason. In principle the laws of collision between molecules in a gas are known. However, most properties of the gas such as temperature, pressure, etc. are dependent only on the average behaviour of large numbers of molecules and are insensitive to the actual details of the collision. Macroscopic laws such as equations of state are deduced using only the laws of conservation of energy and momentum. Some other processes, such as those of heat conduction and diffusion, depend on the actual collisions in a more particular way, but it has been found that even here a satisfactory qualitative theory can be obtained using only one collision parameter. This latter might be, for instance, the size of the molecule or the frequency of collisions. This means that quite crude models can give satisfactory descriptions of macroscopic properties. If this were also true of traffic situations such descriptions of macroscopic behaviour as the flow-concentration relation might well be deducible from crude models in which the exact overtaking mechanism is not completely specified.

The mathematical formulation of this theory is based on an equation which expresses the competition between the wishes of drivers and the constraints exerted on them by other drivers. Suppose that the frequency distribution function which characterizes mathematically the speed of the traffic is $f(x, u, t)$ where $u = \mathrm{d}x/\mathrm{d}t$. The constraints arising from interactions with other vehicles tend to displace this distribution function towards lower speeds.

Now
$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + u\frac{\partial f}{\partial x}.$$

We can divide the term $\mathrm{d}f/\mathrm{d}t$ into two parts. One is the net change in f caused by 'collision' processes in which vehicles, when interacting with vehicles ahead which are moving with slower speeds, either overtake or assume the slower speed. The other is a 'relaxation' process in which the drivers' desire to return to the desired speed distribution function is expressed.

The term due to 'collision' can be written

$$k(\bar{u} - u)(1 - p)f,$$

where p = probability of overtaking,

k = concentration of vehicles $= \int_0^\infty f\,\mathrm{d}u$,

and $\bar{u}$ = average speed $= \int_0^\infty uf\,\mathrm{d}u/k$.

The term due to 'relaxation' can be written

$$\frac{-(f-f_0)}{T},$$

where T = relaxation time,

and f_0 = desired speed distribution function.

Hence the Boltzmann-like equation for $f(x,u,t)$ is a complex integro-differential equation:

$$\frac{\partial f}{\partial t}+u\frac{\partial f}{\partial x} = -(f-f_0)/T+k(\bar{u}-u)(1-p)f.$$

We note that this formulation can deal with the multiple-lane traffic situation since overtaking is allowed for. It has been suggested that T is some function of u.

In general this equation is difficult to solve and we shall mention only one particular case here, namely the stationary solution.

The time-independent $(\partial f/\partial t)=0$), homogeneous $(\partial f/\partial x=0)$ solution for f is

$$f = f_0/[1-kT(1-p)(\bar{u}-u)]. \tag{4.7}$$

This describes a situation in which there is equilibrium between the slowing down of vehicles caused by collision processes and the speeding up of vehicles caused by overtaking. If the dominant relaxation process is that of overtaking the relaxation time T is a function of p which must behave so that $T \sim 0$ for $p=1$ (very low concentrations) and $T \sim \infty$ for $p=0$ (very high concentrations). Hence we can take

$$T = \tau(1-p)/p, \text{ where } \tau \text{ is a constant.}$$

It could also be assumed that p is a function of the concentration, i.e. $p=p(k)$, and to be explicit we could take

$$p = \begin{matrix} (1-k/k_p), & k < k_p \\ 0 \quad , & k \geqslant k_p \end{matrix}$$

where k_p is the limiting concentration on an n-lane road when vehicles can no longer overtake.

Substituting in equation (4.7) for T and p we get

$$\frac{f_0}{f} = 1-\frac{\tau k^3}{k_p(k_p-k)}(\bar{u}-u) = 1-\gamma(\bar{u}-u), \text{ say}$$

i.e.
$$f = \frac{f_0}{(1-\gamma\bar{u})+u}, \tag{4.8}$$

where $(1-\gamma\bar{u})$ is constant for a given concentration k. Then k is determined by normalizing f to the concentration k, so that

$$k = \int_0^\infty \frac{f_0\,\mathrm{d}u}{(1-\gamma\bar{u})+u}. \tag{4.9}$$

Equation (4.9) then gives $(1-\gamma\bar{u})$ and equation (4.8) gives f since f_0 is a known function. All the moments of the distribution can thus be obtained, including the mean speed, and all the relevant hydrodynamical quantities such as q can be computed.

Numerical results for various choices of f_0 are given in Anderson, Herman and Prigogine (1962), together with graphs of flow-concentration curves for the same functions. These show that for low values of the concentration, the relationship is approximately linear, and that for high concentrations, differences in average desired speeds (for any f_0) make little difference to the flow curve.

In general, non-stationary solutions of linearized Boltzmann-like equations are difficult to deal with as it is necessary to solve a non-linear integro-differential equation. It can be done by a perturbation technique which deals with small spatial and temporal deviations from the solution of (4.7). Starting with a perturbation it is possible to use a Fourier-Laplace transform technique to linearize the equation.

The kinetic theory just outlined is primarily applicable to multi-lane high-speed roads under light traffic conditions, so this theory to some extent complements the continuous flow analogy. In this analysis several of the approximations current in kinetic theory have been applied to traffic problems involving interactions. The results are plausible but the methods are rather *ad hoc* in the traffic context, the interactions specified being vague and difficult to relate to the actual behaviour of drivers. The deductive nature of this theory may, therefore, be open to question.

CHAPTER 5

The Fundamental Diagram of Traffic II

One of the basic problems in traffic flow theory concerns the theoretical derivation and experimental verification of the flow-concentration relation or equation of state. Let us consider, therefore, how the fundamental diagram of traffic can be deduced from the various theories and what form it takes in each case.

Consider first how the fundamental diagram can be obtained from microscopic considerations, i.e. by considering the case of two particular vehicles following one another: this has already been described in Chapter 3. A considerable body of opinion supports the contention that there are two kinds of following behaviour, one for dense and one for sparse traffic. This would imply two functional forms for the fundamental diagram with the possibility of a discontinuity, but the boundary location for this has never been specified.

Equations (3.9), (3.12) and (3.13) which follow from this theory are given again here for convenience:

$$q = c(1-k/k_j), \quad \text{where } c \text{ is a flow-rate;} \tag{5.1}$$

$$q = ck\ln(k_j/k), \quad \text{where } c \text{ is a velocity;} \tag{5.2}$$

$$q = u_f k\,\mathrm{e}^{-ck}, \quad \text{where } c \text{ is a spacing.} \tag{5.3}$$

Differentiating q with respect to k and setting $\mathrm{d}q/\mathrm{d}k=0$ gives the value of k for q(max), i.e. the optimum value of the concentration for capacity flow. For the form (5.2) this is k_j/e or 37% of jam concentration. The optimum speed in this case is c. For the form (5.3) the optimum concentration is $1/c$ and the optimum speed u_f/e.

Next let us consider how the fundamental diagram can be deduced from the fluid analogy. If $q=q(k)$ and $u=u(k)$, differentiation of $q=uk$ gives

$$\frac{\mathrm{d}q}{\mathrm{d}k} = k\frac{\mathrm{d}u}{\mathrm{d}k}+u, \tag{5.4}$$

and then the equation

$$\frac{dq}{dk}-u = c$$

expresses the hypothesis that the speed of the shock-wave in the traffic fluid is constant. Substitution into equation (5.4) gives

$$k\frac{du}{dk} = c,$$

and the solution is

$$q = ck\ln(k_j/k). \tag{5.5}$$

Notice that this equation, which was also obtained by Greenberg (1959), is the same as equation (5.2). This will be discussed later.

In addition two experimental forms have been suggested. If we assume a linear relation between mean speed and concentration, as indicated by regression techniques, we get

$$\frac{u}{u_f}+\frac{k}{k_j} = 1,$$

leading to the equation

$$q = ku_f(1-k/k_j). \tag{5.6}$$

This is the form obtained by Greenshields (1935). It is a parabola, and optimum concentration in this case is $k_j/2$ or half jam concentration. The other form which has been suggested is

$$q = \frac{u_f k(k_j-k)^{1/2}}{Au_f k^2+(k_j-k)^{1/2}}. \tag{5.7}$$

This, however, seems somewhat arbitrary.

Finally let us consider how the fundamental diagram may be deduced from a stochastic approach. Suppose, following Haight (1958), that each vehicle travels at a speed x whenever it is free from interference from other vehicles. Let these free speeds have a probability distribution with a probability density function (p.d.f.) $f(x)$, and finite mean $E(x) = u_f$.

The speed x will not be maintained all the time, so let the actual average speed be y, where $y = y(x, k)$; this is really the function $y(x, k, u_f)$ normalized by putting x for x/u_f and y for y/u_f.

The distributions of x and y are difficult to determine but some plausible statistical assumptions can be made. In addition there is the basic concept that $k < k_j$ where $k_j \approx 350$ vehicles per mile for a single lane, if each vehicle is taken as approximately 15 ft in length.

There are three limiting conditions which must be satisfied by the distribution of y. Since a vehicle travels at its free speed when there is no traffic, $y(x,0) = x$. Since every vehicle stops in a jam, $y(x,k_j) = 0$. Also, since a vehicle can stop at any time, $y(0,k) = 0$.

The domain for x is $0 < x < \infty$. Since there exists a maximum average speed as the free speed increases, y must satisfy

$$y(\infty,k) = L(k),$$

where $L(k)$ is dimensionless and $L(0) = \infty$, $L(k_j) = 0$. In other words the probability density function must be such that it is defined in $(0.L)$ and must have the limiting property of approaching a probability density function defined in $(0,\infty)$ as $k \to 0$. This is a well-known property of the Pearson Type I or beta distribution, which is discussed in Chapter 7. Its p.d.f. is

$$Cy^{\alpha-1}(L-y)^{L\beta-1}, \quad 0 < y < L$$

where C is given by

$$C = \frac{1}{B(\alpha, L\beta)\,L^{L\beta+\alpha-1}}$$

and α and β are parameters.

The mean of y is given by

$$\begin{aligned} E(y) &= C\int_0^L y^{\alpha}(L-y)^{L\beta-1}\,dy \\ &= \frac{B(\alpha,+1\,L\beta)\,L^{L\beta+\alpha}}{B(\alpha,L\beta)\,L^{L\beta+\alpha-1}} \\ &= \frac{\alpha L}{\beta L+\alpha}, \quad \text{where as } L \to 0,\ E(y) \to 0. \qquad (5.8) \end{aligned}$$

In dense traffic $k \to k_j$ and $L \to 0$. In light traffic $k \to 0$ as $L \to \infty$, and the Pearson Type I distribution becomes a Type III or gamma distribution. This is also the distribution for x. It has p.d.f.

$$C'x^{\alpha_0}e^{-\beta_0 x}, \quad 0 < x < \infty$$

where $$C' = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}$$

and α_0, β_0 are limiting values of α and β.

The mean of this distribution is given by

$$E(x) = C' \int_0^\infty x^{\alpha_0+1} e^{-\beta_0 x} dx$$
$$= \frac{\beta_0^{\alpha_0} \Gamma(\alpha_0+1)}{\beta_0^{\alpha_0} \Gamma(\alpha_0)}$$
$$= \frac{\alpha_0}{\beta_0}.$$

Hence $u_f = \alpha_0/\beta_0$.

By virtue of equation (5.8) the problem of determining a functional form for the fundamental diagram is reduced to the problem of finding the average speed of an infinitely fast vehicle as a function of traffic density, i.e. of finding $u(k)$. It is possible that α and β might depend on k, but this is not necessarily the case since the convergence of L to zero at maximum congestion guarantees that $u(k_j) = \mathrm{var}(k_j) = 0$.

Two functions satisfying the boundary values $u(k_j) = 0$ and $u(0) = u_f$ have been given by Haight. With the normalization removed these are:

(i) $L = \ln(k_j/k)$. Equation (5.8) then gives

$$u(k) = \frac{\alpha u_f \ln(k_j/k)}{\beta \ln(k_j/k) + \alpha}$$

for some α and β, and

$$q = \frac{\alpha u_f k \ln(k_j/k)}{\beta \ln(k_j/k) + \alpha}. \tag{5.9}$$

(ii) $L = \frac{\pi(k_j - k)}{2k}$. Equation (5.8) then gives

$$u(k) = \frac{\alpha u_f \pi(k_j - k)}{\beta \pi k_j + (2\alpha - \beta\pi) k}$$

for some α and β, and

$$q = \frac{\alpha u_f \pi k(k_j - k)}{\beta \pi k_j + (2\alpha - \beta\pi) k}. \tag{5.10}$$

Let us now consider critically the different forms for the fundamental diagram as given by the above eight equations. None of these forms has been fully proven. Only the two experimental forms (5.6) and (5.7), and the two derived from statistical considerations (5.9) and (5.10) satisfy all the boundary conditions. Equation (5.6) is obtained by assuming a linear relation between q and k, which is only approximately true in practice. Equation (5.2) gives a better fit to experimental data, but does not satisfy all the boundary conditions.

Notice that equation (5.2) is obtained from a consideration of two vehicles following one another. In this case c is the parameter defining the sensitivity of the response in the reciprocal spacing model, and is sometimes termed the 'characteristic speed'. Equation (5.5), on the other hand, is obtained from macroscopic considerations. Here c is the optimum speed measured from experiments involving many vehicles on a road, and has been found to be of the order of 20 m.p.h. The fact that it is possible to relate the characteristic speed of the microscopic model with the optimum average speed of the macroscopic model is quite remarkable.

CHAPTER 6

From Traffic Signals to Traffic Cybernetics

6.1 Introduction

Before dealing with two other approaches to problems of traffic flow – stochastic theories and simulation – we shall discuss an important topic, namely that of traffic signals.

One possible improvement which can be made to an uncontrolled (priority) intersection is the installation of traffic signals. This requires only the addition of the necessary signal equipment, and a change in layout is not strictly necessary, though such a change is often made. A general description of the mode of operation of such signals is given in §6.2. The delay experienced at fixed-cycle signals is discussed in §6.3, and in §6.4 the related topic of the formulae to be used for the optimum setting of such signals. Pedestrian crossings are considered in §6.5, and in §6.6 we take a look into the future.

The geometry of road layout, including such things as roundabouts, will not be considered here. These latter are, of course, involved in problems of traffic flow in a similar way to signals, but they are perhaps more appropriately dealt with in books on traffic engineering.

6.2 Traffic Signals

Traffic signals of the kind we know today were first installed in Britain in 1926, although manually operated semaphores had been in use since 1868. They are of two main types. The simplest kind are automatic fixed-cycle signals giving alternate stop and go periods of a predetermined length. The sequence in Britain is red, red and amber together, green, amber; the amber is standardized at 3 sec and the red-amber at 2 sec. The more flexible modern kind, which, for the most part, have replaced the fixed-cycle type in Britain, are vehicle-actuated, i.e. the duration of the green time varies according to traffic demands. On roads with signal-controlled intersections at short intervals a linked system is usually used. This is either a limited or flexible progressive system in which a vehicle passing one green signal is faced with successive greens (or green wave) if travelling at a certain speed, usually about 20 m.p.h. for a cycle-time of 1 min. This type of system helps the flow and tends to minimize the delay. Altogether there are about 4,000 sets of signals of all kinds in Britain.

Further details about vehicle-actuated signals are as follows. There is either a fixed minimum green period, which is sufficient to allow all the vehicles waiting between the detector and the stop-line to clear the junction, or sometimes a variable minimum green period which varies between 8 and 15 sec according to the number of vehicles waiting. Following the minimum green is a vehicle-extension period, in which any vehicle crossing the detector causes an extension of the green period to enable it to pass about 10 to 20 ft beyond the stop line before the light changes, the extension time varying according to the speed of the detected vehicle. The green-time extension starts when the vehicle crosses the detector pads, which are usually situated about 90 to 130 ft from the stop line. When the interval between vehicles crossing the detector becomes greater than the vehicle-extension period, a change of phase occurs if required. A maximum green period is set so that when there is a continuous stream of traffic on the road which has right of way, long delays to vehicles on the opposite arm are avoided. At some junctions carrying very dense traffic the green period may run to a maximum on all phases giving, in effect, fixed time working. An important variation is an early cut-off of 1 or 2 sec or a late start of the green period on the opposite arm. These all-red periods help to clear the intersection of turning vehicles, give pedestrians a few extra seconds in which to cross and lessen the risk arising from vehicles that attempt to cross during the amber or the beginning of the red periods. It has been shown that this arrangement has helped to reduce accidents due to cross-road collisions.

The Ministry of Transport is now (1964) considering an improved form of variable maximum period to suit conditions on roads (especially multi-lane roads) carrying dense traffic, where signals often act virtually as fixed-cycle signals because of the heavy flow.

In ordinary linked systems of traffic signals, vehicles sometimes arrive at the next intersection too soon, have to stop and, when starting up, delay other vehicles which are travelling at the correct speed.

In some linked systems the correct speed may vary over different parts of the system and it may be difficult for a driver to keep at the correct speed for progression throughout the whole system, although notices of this speed are always posted. In addition vehicles entering from side-roads are often out of phase with the progression. There are, therefore, some disadvantages inherent in this sort of scheme.

Linked systems require master controllers to control the individual signals. This idea could be extended by joining up such linked systems to form a network controlled centrally by an electronic computer. This is discussed in §6.6.

6.3 Formula for the Delay at a Signal-Controlled Intersection

It has been estimated that in Britain delays at traffic signals amount to about 100 million vehicle-hours each year. The potential financial saving in the correct setting of signals for minimum delay is therefore considerable. We shall consider here only two expressions for the average delay at a signal-controlled intersection; consideration of other work on this problem is deferred until Chapter 8.

The formula due to Webster (1958) for the delay is an empirical formula, and applies to vehicles arriving at random at fixed-cycle traffic signals. It was obtained by simulation of the situation on an electronic computer, assuming random arrivals for the vehicles. Theoretical methods for calculating the delay at fixed-cycle signals were not very successful at the time, though some attempts have been made since which agree substantially with the results obtained from using Webster's formula. This formula gives the average delay per vehicle for a single approach to an intersection. It is the difference between the average journey-time (in seconds) through the intersection and the time for a run which is not stopped or slowed down by signals. Webster's formula is

$$d = \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} - 0{\cdot}65\left(\frac{c}{q^2}\right)^{1/3} x^{(2+5\lambda)}, \tag{6.1}$$

where c = cycle-time, i.e. one complete sequence of phases;

$\lambda = g/c$, i.e. proportion of cycle-time 'effectively green';

q = flow, i.e. average number of vehicles/second;

s = saturation flow;

$x = q/\lambda s$, i.e. degree of saturation.

The above terms need some explanation. 'Effective green' time is (green + amber − 2) sec, the 2 sec being an allowance for delay in starting. The capacity of an intersection depends on the saturation flows of the individual approaches, and on the amount of lost time in the cycle. The lost time in a cycle is a somewhat variable quantity: it may be anything from 0 to 7 sec, with an average of about 2 sec. Saturation flow is the flow which would be obtained if there were an endless queue of vehicles and they were given a continuous green. The degree of saturation is the ratio of actual flow to the maximum flow which can pass through the intersection. Saturation flow has been measured under controlled conditions, by Charlesworth and Webster (1958), and the results have shown that it varies with approach road width, amount of turning traffic and the number of parked vehicles.

Charlesworth *et al.* gave the figures reproduced in Table 6.1, which shows peak-hour saturation flow in terms of passenger-car units/hour, with no turning traffic and no two-wheel or parked vehicles.

TABLE 6.1

Lane width w ft	Saturation flow p.c.u./h
10	1,850
11	1,875
12	1,900
13	1,950
14	2,075
15	2,250
16	2,475
17	2,700
$17 < w < 60$	$s = 160w$

Let us now look at equation (6.1). The first term represents the delay for uniform rate of flow, the second term corrects for the randomness of the arrivals and the third term is an empirical correction which amounts to only about 10% of the total delay. If d and c are measured in seconds, q and s are in vehicles/second. An approximation to (6.1) is usually used in practice, thus:

$$d = \frac{9}{10}\left[\frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)}\right]. \tag{6.2}$$

Formula (6.1) has been tabulated by Webster over a range of values of the parameters as follows:

$$\begin{aligned} s &= 900(300)1800,2400,3600 \\ c &= 30(5)60(10)120 \\ q &= 50(25)1200 \\ g &= 10(5)100. \end{aligned}$$

As an example of Webster's formula consider a flow of 600 vehicles/h and signal settings of green 29 sec, amber 3 sec, cycle time 60 sec. On average 15 vehicles are discharged in a fully-saturated green period, i.e. 15 vehicles are discharged in $[(29+3)-2]=30$ sec.

Hence
$$\begin{aligned} s &= 1{,}800 \text{ vehicles/h} \\ c &= 60 \text{ sec} \\ \lambda &= g/c = 30/60 = 1/2 \\ x &= q/\lambda s = 600/900 = 2/3 \end{aligned}$$

whence
$$\begin{aligned} d &= \frac{9}{10}\left[\frac{60(1/2)^2}{2(2/3)}+\frac{4/9}{(2/3)(1/6)}\right] \\ &= \frac{9}{10}[11{\cdot}25+4{\cdot}00] \\ &= 13{\cdot}72 \text{ sec.} \end{aligned}$$

Webster also included a formula for the average queue-length, which is the queue-length at the beginning of a green phase and is usually the maximum queue in the cycle. It is given by

$$\operatorname{Max}\left[N = q\left(\frac{r}{2}+d\right),\quad N = qr\right],$$

where r = red time,

q = flow,

d = average delay/vehicle.

This underestimates the queue-length by 5–10% since it assumes that vehicles do not join the queue until they have reached the stop line. A more correct formula is

$$\operatorname{Max}\left[N = q\left(\frac{r}{2}+d\right)\left(1+\frac{qj}{av}\right),\quad N = qr\left(1+\frac{qj}{av}\right)\right],$$

where j = average spacing of vehicles in the queue,

a = number of lanes,

v = free running speed of the traffic.

For light flows, the average queue at the beginning of a green is equal to the number of vehicles arriving during the preceding red, and hence follows a Poisson distribution. For heavier flows, however, the distribution has a considerably longer tail. The point at which a departure from the Poisson distribution occurs is that at which the majority of the cycles are saturated.

The above formulae for the delay and the queue-length apply also to vehicle-actuated signals running to a maximum, but in other cases the delay is very difficult to predict.

Beckman, McGuire and Winsten (1956) have also studied signalized intersections, and have developed the following three formulae for the mean delays for all vehicles queueing during the red phase, the green phase and the complete cycle respectively. If these quantities are denoted by $E(t_r)$, $E(t_g)$ and $E(t_c)$, where E denotes the 'expected' or mean value, then

$$E(t_r) = r\left[E(N_r)+\frac{q}{2}(r+1)\right];$$

$$E(t_g) = \frac{1}{2(1-q)}E[N_g^2-N_r^2-(2q-1)(N_g-N_r)];$$

$$E(t_c) = \frac{r}{(1-q)}\left[E(N_r)+\frac{q}{2}(r+1)\right]; \tag{6.3}$$

where q = arrival rate,

r, g, c = red, green and cycle-times,

$E(N_r)$, $E(N_g)$ = mean queue-lengths at start of red, green.

Dividing equation (6.3) by $q(r+g)=qc$, the mean delay/vehicle is found to be

$$d = \frac{r}{(1-q)(r+g)}\left[\frac{E(N_r)}{q}+\frac{(r+1)}{2}\right]. \tag{6.4}$$

All the quantities in (6.4) with the exception of $E(N_r)$ are known, and the latter can be generated by a Markov chain technique.

A number of other authors, among them Haight (1959), have given various treatments of this subject, but not all have produced results which can be used in practice.

Webster and Wardrop (1962) have made some comparisons between priority and signalized intersections, particularly with regard to capacity. They have come to the conclusion that a number of benefits accrue from the installation of signals. Firstly, saturation flow increases on both main road and minor road, and in all cases more capacity is afforded. If delay-flow curves are plotted for Webster's formula of the delay at traffic signals, and for the formula due to Tanner (1962) of the delay at an uncontrolled intersection (for this see Chapter 8), the point at which the curves cross will determine the point at which signals should be installed.

6.4 Optimum Settings for Fixed-Cycle Signals

A number of authors have dealt with the problem of optimum settings for the phases of fixed-cycle traffic signals, including Webster (1958) and Uematu (1958).

The earliest suggestion for the setting of traffic signals was that the green phases should be in proportion to the corresponding ratios of flow to saturation flow, assuming that these ratios are the same for all arms of the same phase. This can be shown to give approximately the best division of the cycle-time. If, for example, the values of flow and saturation flow for the N–S phase of a simple intersection are 600 v.p.h. and 1,800 v.p.h. respectively, and the corresponding values for the E–W phase are 300 v.p.h. and 1,800 v.p.h., the ratio of green times is given by

$$g_{\mathrm{N-S}}:g_{\mathrm{E-W}} = 600/1{,}800:300/1{,}800 = 2:1.$$

In general the flows on the different approaches belonging to the same phase will not be equal, although the corresponding green will be the same. It can be shown that each phase is best represented by the approach having the highest ratio of flow to saturation flow. Let these ratios be called $y_{\mathrm{N-S}}$, $y_{\mathrm{E-W}}$.

Webster used his formula (6.2) to calculate delays for various values of the cycle-times and ratios of effective green times, taking the lost time, L, as 10 sec. The best ratio for the effective green times, i.e. the one giving minimum delay, was found to be between 1·88 and 2·17 over a range of cycle-times of 35 to 80 sec. The values for the ratio obtained in this way were found to be within a few per cent of the values obtained from the crude method outlined above. For the case quoted, the optimum ratio was 2·07 as against 2·00 for the crude method.

The problem has also been treated mathematically. It was found that, for a 2-phase intersection where the ratio $y_{\mathrm{N-S}}/y_{\mathrm{E-W}} > 1$, the effective green ratio should be rather less than that given by the y values, but in most cases the difference was negligible.

In practice, therefore, it is sufficient to set the effective green times in proportion to the y values, particularly as some of the variables, such as lost time, are only approximate anyway. Then the controller setting of green time is obtained from

controller setting for green = effective green + lost time − amber.

The remaining problem is to find the optimum cycle-time, assuming that the green times are in the ratio of the y values, as just described. For this purpose Webster used equation (6.2) modified to cover the general case of an intersection having n phases instead of two. If this modified equation is differentiated with respect to the cycle time and set equal to zero, a formula giving the value of c for minimum delay can be obtained. This formula is too complicated for practical purposes and a simple approximation is sufficient. This is

$$c = \frac{1{\cdot}5L+5}{1-Y}\,\mathrm{sec},$$

where $Y = \sum y_i$ over the whole intersection,

and L = total lost time/cycle.

The latter quantity is $(nl+R)$ where n is the number of phases, l is the average lost time/phase and R is the time during each cycle when every signal shows red (or red-amber) together.

In order to set fixed-cycle signals the procedure is therefore as follows:

(i) Estimate the flow q and saturation flows s for each arm of the intersection.

(ii) Evaluate the ratio q/s for each arm and put $y = \max(q/s)$ for each phase.

(iii) Set $Y = \sum y$.

(iv) Estimate the lost time R due to all-red or red-amber periods for pedestrians etc.

(v) Calculate the optimum cycle-time from the formula

$$c_0 = \frac{1{\cdot}5L+5}{1-Y}.$$

(vi) Calculate the effective green times from

$$g_i = \frac{y_i}{Y}(c_0-L).$$

(vii) Calculate controller settings for green from

$$g = g_i+L-A$$

where A = amber time (usually 3 sec).

As an example suppose that measurements of flow and saturation flow at a 2-phase intersection are as follows

	N	S	E	W
Flow q	600	400	750	1,200
Saturation Flow s	2,400	2,000	3,000	3,000

Suppose also that there is a starting delay of 2 sec at each phase, and that there is an all-red period of 4 sec between each change of phase. The y values are then

$$y_{\text{N-S}} = \max\left(\frac{600}{2{,}400}, \frac{400}{2{,}000}\right) = 0{\cdot}25,$$

$$y_{\text{E-W}} = \max\left(\frac{750}{3{,}000}, \frac{1{,}200}{3{,}000}\right) = 0{\cdot}40,$$

and $$L = 4(2)+2(2) = 12 \text{ sec.}$$

Then $$c = \frac{1{\cdot}5(12)+5}{1-0{\cdot}65} \approx 66 \text{ sec.}$$

The total effective green time is $(66-12) = 54$ sec, and this must be divided in the ratio 0·25/0·40 giving

$$g_{\text{N-S}} = 21 \text{ sec}, \quad g_{\text{E-W}} = 33 \text{ sec.}$$

The controller settings are then

$$\text{N–S}: (21+2-3) = 20 \text{ sec green} + 3 \text{ sec amber},$$

$$\text{E–W}: (33+2-3) = 32 \text{ sec green} + 3 \text{ sec amber}.$$

Note that under conditions of very light traffic the optimum cycle times as found by the procedure just described might be very short. Having regard to practical and safety considerations, the cycle-time should not normally be less than about 25 sec.

Uematu has used a different approach. He suggested that the lengths of the phases should be determined by the lengths of the queues on the two approaches. A 'random walk' concept was used to describe the lengths of the queues X_n and Y_n on the two approaches for the nth cycle C_n. The fundamental equations are then

$$X_n = (X_{n-1}+x_n-qG_X)+x'_n, \quad n = 1, 2, \ldots;$$

$$Y_n = (Y_{n-1}+y_n-qG_Y)+y'_n, \quad n = 1, 2, \ldots;$$

and $$C_n = G_X+G_Y;$$

where $C_n = n$th cycle,

X_n, Y_n = lengths of queues in C_n,

X_{n-1}, Y_{n-1} = lengths of queues remaining from previous cycle C_{n-1},

x_n, y_n = numbers of arrivals during green,

x'_n, y'_n = numbers of arrivals during red,

q = rate of departure in vehicles/sec,

G_X, G_Y = lengths of green phases in seconds.

If $X_n^{(\max)}$ and $Y_n^{(\max)}$ represent the maximum desired queue-lengths on the two approaches, G_X and G_Y must be chosen so as to minimize the probabilities of the queue-lengths exceeding these maxima. Uematu formulated the transition probabilities which describe the system for any cycle-length, and derived solutions for some special cases.

Further discussion of probabilistic models for signalized intersections is deferred until Chapter 8.

6.5 Pedestrian Crossings

There are at present three types of pedestrian crossings in Britain, viz. ordinary zebra crossings, push-button signals with zebras and pandas which consist of more complicated push-button systems incorporating a traffic signal for vehicles together with pedestrian signals and push-buttons. The first two types date in their present form from 1951 and are well known. They are now in use all over the world. The third type was first installed in 1961 and has turned out to be somewhat unpopular as it is difficult for motorists to differentiate between the flashing and pulsating lights which are used. The vehicle signals in this case are similar to those used with standard traffic lights but there is no green signal. The pedestrian signal consists of an indicator with the word CROSS in white on a black ground, which is operated when the pedestrian pushes a button. The indicator reading WAIT is illuminated at appropriate times and shows as white on a blue ground. A buzzer is included on some installations as an indication to the blind of the right of way.

In September 1963 a new pedestrian control experiment in Britain was launched by the Ministry of Transport. Section 8(v) of the Road Traffic and Roads Improvement Act 1960 makes it clear that under Section 34 of the Road Traffic Act 1960 the Minister of Transport can control pedestrians in the London Traffic Area by means of traffic regulations. The experiment, which is still in progress at the time of writing (1965), is an attempt to secure greater road safety by some combined regulation of the motorist and the pedestrian. It is being carried out, for 2 years in the first instance, at three sites in London, each of which is considered to be a typical inner suburban shopping street area and each with a bad accident record.

Along these stretches of road it is an offence for pedestrians to cross the road or walk in the carriageway except at controlled crossing places, or to step off the pavement at a crossing place except when the special light signal for pedestrians shows 'CROSS'. The traffic signals are so phased that no turning movements of traffic interfere with the safety of pedestrians crossing at the right time.

The areas are all clearly marked CONTROLLED AREA and include instructions to pedestrians. Motorists merely comply with traffic lights in the usual way. A novel feature of the scheme is the inclusion of an illuminated symbolic figure of a pedestrian whose arms and legs move in a representation of walking when the pedestrian is allowed to cross. A pedestrian registers his demand through a push-button, but vehicular traffic is not required to stop until a set period has expired.

It might be of some interest to say a few words about the legal aspects of this type of experiment. For some time it has been an offence for pedestrians to cross the road if signalled to stop by a traffic policeman. However, this is the first time they have been prohibited by regulations from crossing the road or walking in the carriageway where they like. It has already been recognized that the success of such an experiment depends a great deal on public co-operation, and one way that was used to achieve this was a massive propaganda campaign which included some 3,000 display posters, 6,000 leaflets and a film.

There are some doubts, however, whether such a scheme can be made to work in practice as a regular routine, due to such factors as the attitude of the public, the powers of the police and the attitudes of the courts. In New York similar schemes have been tried since 1958 and in the year following the introduction of the new law there was a drop of 6% in pedestrian fatalities and further drop of 21% in 1960. However, the prognosis for the London experiment may not be as good because the population as a whole in Britain is perhaps rather less pro-motorist than in the U.S.A. In addition, the actual power of the the police to enforce regulations is small, there being no power to arrest, nor even power to make the offender give his name and address, although pedestrians may at first not realize this. If the courts were to impose a series of derisory fines for pedestrian contravention of the regulations this would be harmful: the maximum penalty which can be incurred is, in fact, £20.

The actual results of the experiment will not be known for some time, but an optimistic view is held by at least some of the people concerned. The police view is that accident rates must drop where such schemes operate and shopkeepers seem to be generally in favour. Although it may seem odd at first sight, opposition seems to be coming from the motorist rather than from the pedestrian who is generally in favour of the scheme. This is probably because, although the linked system is designed to reduce traffic delays to a minimum, the special phases for pedestrians to cross with complete protection inevitably mean some increase in delay to vehicles. The unfortunate effect in some cases is that some of the traffic has been diverted to other routes on parallel roads. This will make comparisons of some of the factors in before-and-after experiments difficult.

It is a well-known fact that in Britain approximately 90% of all accidents to pedestrians occur in built-up areas. About 80% of these pedestrians are injured whilst crossing the road. In 1962 no less than 2,000 people were killed and 60,000 injured on urban roads. However, the average rate of increase of pedestrian casualties, which is 2% p.a. over the last 10 years, is still considerably less than the rate of increase in the amount of traffic, which is 7·4% p.a.

The number of pedestrians using a crossing varies with a number of factors, including the site of the crossing, the flow of vehicles, the time of day and so on. At some places the number of pedestrians using the crossing expressed as a percentage of all those crossing within 20 yd on either side is as high as 80%, the proportion for women being higher than that for men. The risk to pedestrians on a crossing is about 30% of that off it. Data on pedestrian casualties have shown that the safest crossing points are traffic lights, and the next safest zebra crossings. Busy crossings are safer than the less busy, presumably because the hazards are well known and drivers take special care. Crossing at a distance of 50 yd or so from lights or zebras is especially risky, presumably because at these points drivers have their attention fixed on the hazard ahead and pay less attention to the immediate curbs.

Before-and-after studies suggest that refuges and zebra crossings do affect the places at which pedestrians cross. In the case of zebras, counts of numbers of pedestrians crossing the road within 50 yd of a zebra, and of those crossing outside the 50 yd limit, suggest that the effect extends to perhaps 50 yd on either side. It has also been shown that there is an increase of about 1% in the proportion of all pedestrians who use crossings for every increase of 24 vehicles/h in the flow of traffic using the road, over a range of flow of 200 to 1,000 vehicles/h.

Measurements have been made of delays to both vehicles and pedestrians at both zebras and pedestrian-operated signals. Comparisons have been made of the delays experienced at a zebra and at a signal installation on the same road, and before-and-after studies have been carried out when zebras have been replaced by pedestrian-operated signals.

Journey-times of vehicles over a stretch of road on either side of a zebra can be measured, and simultaneous counts taken of the numbers of pedestrians crossing. A regression equation

$$j = a+bp,$$

where j = average journey time in seconds

p = number of pedestrians/h at the zebra,

can be fitted to the data, and actual experiments have suggested that the journey-times of vehicles passing the zebra increase approximately linearly with the number of pedestrians crossing, the journey time increasing by 1 sec for an increase of about 170 to 250 pedestrians/h.

As far as delays to vehicles at signals are concerned, phases are usually set so as to limit the time available to pedestrians. Hence the delay to vehicles is not increased by an increase in pedestrian flow once the signal changes have reached maximum frequency. Up to this point delays increase approximately linearly with the number of pedestrians. We then have

$$y = \begin{matrix} a+bp, & \text{for small } p \\ c, & \text{for large } p. \end{matrix}$$

Results in this case are, of course, affected by the setting of the signals and the readiness or otherwise of pedestrians to press the button.

Comparison shows that zebras cause less delay to vehicles than signals at low pedestrian flows, but that the reverse is true for high flows, the critical level being about 1,000 to 1,500 pedestrians/h. Average figures for delays to pedestrians are between 3 and 4·5 sec at a zebra, and between 7 and 10 sec at signals.

In the case of the pedestrian control experiment mentioned earlier, before-and-after studies are being carried out by the Road Research Laboratory. In this study counts are being kept of traffic flows in both directions and pedestrian delays at crossings. In addition, estimates of delay to vehicles caused by pedestrians, traffic and traffic lights have been made, together with counts of flows on parallel roads. The results of this trial are not yet available.

6.6 The Future

Simple rules can sometimes be very effective in handling complex situations. For example, the rule 'drive on the left' prevented chaos on the roads of Britain in the early days of motoring. The equally simple rule 'never turn across a stream of traffic in cities' might ease the traffic flow in the centre of cities to a considerable extent. If, instead of holding up traffic as in Fig. 6.1(a) drivers had to keep turning left until they could cross the traffic stream at right angles as in Figure 6.1(b), junctions controlled by traffic lights would handle a greater capacity than they do at present. This has already been tried in London at a number of sites, with some success. Increased flows have certainly resulted, and bus drivers, for instance, have benefited

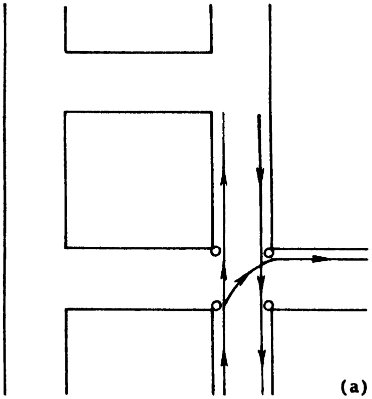

FIG. 6.1(a). Diagram of a controlled intersection with no restriction on R-turning traffic.

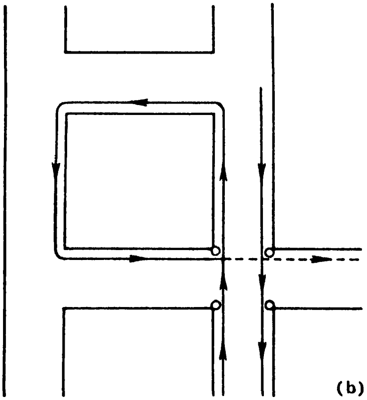

FIG. 6.1(b). Diagram of a controlled intersection with a ban on R-turning traffic.

considerably from the consequent reduction in delays due to this, and also to increased one-way working. Necessary re-routing has, of course, caused some inconvenience to passengers by the lengthening of some journeys and the re-siting of stops, but on the whole their gain has been greater than their loss.

The Ministry of Transport has in the past few years already tried to decrease delays in Central London in a number of ways – by using traffic signals more efficiently in the way described above, by diverting traffic onto alternative routes, by one-way working and by tidal flow. This last consists of the technique of lane-switching whereby more than half the lanes in a multiple-lane road are used to carry traffic in one direction during the morning and evening rush hours.

As early as 1920 it was suggested that a planned traffic flow could be achieved by applying fixed-cycle signals to random traffic flow. Next came the idea of co-ordinated control of batches of traffic, so as to achieve a traffic flow over several intersections without stops. Fixed-cycle signals gave way to traffic-actuated signals to provide a more flexible system and to avoid unnecessary delays. Under present conditions the need for control based on traffic intensity has given way, to some extent, to the need for a constant flow, and co-ordination has become a necessity.

Ways of co-ordinating whole traffic systems are being sought: this is called traffic cybernetics. The so-called 'progressing system' is usually aimed at. One-way progression for every driver is apparently the most sensible system because then the speed of the vehicles is co-ordinated with the progression of the green light. This sort of system is suitable for tidal flow. It can be extended to two-way progression which is more in accord with practical situations. Here the distances between crossing streets is important and attention should be paid to this when building new towns or rebuilding old ones. These ideas gives rise to the idea of 'dynamic town-building'.

A fundamentally integrating element is the speed indicator or pre-signal to drivers on the approaches to the linked system which enables every driver to accommodate himself to the correct speed. These are usually called 'intermediate funnels' or 'Düsseldorf funnels', after the name of the town in Germany where they were first installed in 1954. Many are to be found in Germany and Switzerland.

The basic idea of dynamic traffic cybernetics is the complete space-time co-ordination of all the processes of movement in central city areas, in order to achieve maximum traffic capacity with the smoothest possible traffic flow. The aim is to use constant electronic control to guide the whole mass of traffic both by means of routing and signal control so that jams cannot occur at peak periods. Under such a scheme the traffic would be collected in preliminary areas, and only enough allowed to enter the critical areas to keep them just below saturation point. A digital computer would do a rapid analysis of traffic data (with scans every 2 sec, say) so that optimum flow would be assured.

Area control of traffic can have a number of different aims, some of which, unfortunately, conflict. One important aim is the reduction of total delay, since this has an economic justification. Another is the increase of capacity, and this applies particularly at peak periods. This conflicts with the minimization of the total delay to the main stream at a junction. Another aim is the maximum utilization of the network under peak conditions by the diversion of the traffic from overloaded routes to alternative routes with spare capacity. This is called route recommendation and control.

There are a number of ways in which a computer can be used in network control. In the first scheme knowledge of average flows and journey speeds on certain sections of the network and at various times throughout the day could be used to construct master programs. This is an extension of present practice. These could then be applied to the system either on a time basis, or on the results of sample traffic counts taken at a number of strategic points within the system. Obtaining master programs to minimize the overall delay to vehicles would, of course, be quite a difficult task. Several ways of linking simple systems are known already and simulations on electronic computers could be carried out to determine the best arrangements of signal settings as far as total delay is concerned, but it has been shown in the United States that this is tedious and slow. Analogue computers might also be used.

The second scheme is based on the principle that the area should be linked by signals in such a way that the network as a whole has the highest attainable capacity. This might involve making certain streets one-way, where linking is known to be more efficient, and bringing into force other restrictions with the object of improving overall capacity quite independently of existing traffic movements and volumes. Some of these ideas have already been applied to selected areas in Central London.

Thirdly, computers could be used to control a series of moves in the manner of a game of chess. In this case there is no master plan for the controller, but at fixed intervals of time a scan is carried out and a decision made by the central computer as to whether a particular signal should be changed or not. The decision would be based on information obtained from detectors installed at every intersection of the network, these detectors being of an advanced type capable of measuring both the flow and either speed or concentration. The information thus obtained would be fed continuously to the computer. Thus at each instant the computer would record how much traffic was on each section of the network and its average speed. A prediction, with certain assumptions, could then be made for the next few minutes for each of several different signal settings. The setting to be used could then be chosen by optimizing the total delay to vehicles, or according to some other criterion. This method has the advantage that it could automatically take account of temporary bursts of traffic – the emptying of a cinema, for example. Such a program would be tedious to write but it is possible. A large, fast computer, such as Atlas or KDF 9 would, however, be absolutely essential.

We note that in Britain existing linked signals are vehicle-actuated, not fixed-time, so that area control would make a less marked improvement than in the United States or Canada, for instance, where less flexible fixed-time systems are more common.

The electronic equipment to be used in this kind of traffic control would have to deal automatically with the normal range of conditions to be expected, and should monitor the operation of the individual controllers and the traffic flows so that accidents, road-works, and so on could be detected. An alarm could then bring in, say, manual control or fixed-cycle working. British vehicle-actuated systems already provide most of the facilities needed for on-site equipment, but ancillary equipment connected with the computer may need development.

Vehicle detectors would provide much of the basic traffic information on traffic flows. Recently a more sensitive type of detector has been developed, and several variants are now available. These detect the mass of metal, register whether it is moving, give an idea of its speed and count the length of the queue.

The detectors would be connected by telephone lines to input devices connected to a computer, and decisions would be returned from the computer to the traffic control boxes. As this would occupy only about 200 msec out of every second the rest of the computer time could be spent on other jobs. It is envisaged that approach detectors would have to be placed about 300 ft from the stop-line in order to give 8 to 10 sec warning of the arrival of a vehicle. The data transmission systems used for transmitting data to and from the computer and the signals are already developed, as are the computers themselves.

At a Traffic Engineering Exhibition held in London in July 1963 a comprehensive demonstration was held of Ferranti traffic control equipment based on the use of an Argus 104 computer. The Argus was demonstrated working in conjunction with a model of a traffic network in a large city. Linked to the model, on-line, Argus performed a variety of jobs including the sensing of model vehicles approaching the cross-roads, counting the vehicles and controlling the signals in accordance with the queue-lengths, subject to a time-limit, and computing the speeds of the vehicles for printing out if required. As the Argus was equipped with an interrupt system it could easily be switched to an additional task, such as an engineering calculation connected with traffic flow, in the spare milliseconds between the phases of the traffic control routine.

In Amsterdam one plan has involved main-road traffic being centrally controlled. This only works when traffic on the minor-roads is light. In Canada, the city of Toronto is working on an ambitious scheme to bring its road traffic under computer control. To date, the scheme has resulted in an 11% reduction in the overall average delay due to queueing at peak periods, and 25% in the average delay due to queueing at the morning peak. In addition there has been a 20% increase in traffic volume for a given fixed amount of delay.

In Britain an experiment in an area of West London could eventually lead to the use of a central computer, a measure intended to postpone (at least) complete traffic paralysis. In the Ministry of Transport's £425,000 experiment it is expected that four square miles of West London may come under computer control, and this experiment is the only one in the world on the same scale as that of Toronto. The area to be covered is bounded by Knightsbridge, Hammersmith, the Kensington Road and the Thames.

There are extra difficulties in the scheme for London as compared to that of Toronto. Firstly, the gridiron layout of roads in Canadian cities simplifies matters there. Secondly, in the case of London an equally congested area lies on the perimeter of the experimental area, and congestion may only be transferred, not eliminated.

The London experiment started in July 1961 with an enormous traffic survey, the largest and most comprehensive traffic study in the country. It is divided into three phases. Phase 1 consisted of a comprehensive inventory of current traffic volumes, origin-destination surveys and so on. Phase 2 includes forecasts of future patterns in traffic movements. Phase 3 will consist of an analysis of the adequacy (or otherwise) of existing roads and the evaluation of various long-range alternative proposals. The question will have to be decided whether a digital computer will be needed or whether an extension of the existing analogue system will be sufficient. The tremendous advantage of a digital computer would be the ease with which a traffic control program could be changed. Closed circuit television is also envisaged in case an emergency – a bottleneck outside the perimeter, for example – necessitates the overruling of the electronic system.

Professor Buchanan, in his report Traffic in Towns (1963) has stated that, in his opinion, the paralysis which is threatening our major cities will only be delayed, and not cured, by such measures as those outlined above, and that more imaginative and more drastic measures will be necessary if chaos is not to result. This may well be so, but further discussion of this matter would be out of place here.

CHAPTER 7

Statistical Distributions and Queueing Theory

7.1 Random Distributions

Road traffic can be considered either as a distribution in space or as a distribution in time. In the first case we have a certain length of road (considered as a single lane) and the vehicles at any given moment in time are distributed along it. In the second case we have a certain fixed position on the road and the vehicles pass this position at various time intervals. The succession of spacings (in the first case) or time-intervals (in the second) form distributions of vehicles in space or in time. These distributions may take, or approximate to, various mathematical forms. The space or time intervals could, for example, at least in theory, all be exactly equal. More likely, the time and distance gaps between vehicles will be distributed in a more haphazard manner with large and small gaps occurring in different proportions. These distributions may sometimes be 'random', where the word is here used in its stochastic sense and not as a synonym for vaguely haphazard.

A series of events (or distribution) is defined as random in this sense when:

(a) each event, for example the moment of arrival of a vehicle at a given point, is completely independent of any other event; and
(b) equal intervals of time (or space) are equally likely to contain equal numbers of events (or vehicles).

Under these circumstances two distributions are relevant. One of these distributions gives the probability of any given number of vehicles passing a point during a specified time, or of being in a specified length of road: this is called the Poisson distribution. The other gives the probability of getting a gap (time or distance) of a given length in a specified time or over a specified distance: this is called the negative exponential distribution. These distributions are now derived.

The Poisson distribution can be derived in a number of ways, of which the following is perhaps the most instructive in this context. The derivation is couched in terms of the time t rather than the distance x, but the distribution applies equally to both cases.

Let events, i.e. arrivals of vehicles, occur in such a way that:

(i) the probability of an event in $(t, t+\delta t)$ is $\lambda\delta t+o(\delta t)$;

(ii) the probability of two or more events in $(t, t+\delta t)$ is $o(\delta t)$;

(iii) the number of events in $(t, t+\delta t)$ is independent of what has happened in $(0, t)$.

Here λ is a constant and $o(\delta t)$ indicates a small-order quantity which can be neglected. Thus (i) and (ii) taken together mean that the probability of an event occurring in a small time-interval is proportional to the length of that time-interval, and that the probability of two or more events occurring in this interval can be neglected. In the case of very dense traffic drivers do not act independently of one another and both (ii) and (iii) break down.

Then if $P_k(t)$ is the probability of k events occurring in time t, the probability of k events occurring in time $(t+\delta t)$ can arise in two ways. Either k events occur in time $(0, t)$ and no event in $(t, t+\delta t)$, or $(k-1)$ events occur in $(0, t)$ and one event in $(t, t+\delta t)$. All other combinations which make k events altogether are small enough to be neglected.

Hence

$$P_k(t+\delta t) = P_k(t)[1-\lambda\delta t]+P_{k-1}(t)\,\lambda\delta t+o(\delta t),\ k = 1, 2, \ldots.$$

and $P_0(t+\delta t) = P_0(t)[1-\lambda\delta t]+o(\delta t)$.

Subtracting $P_k(t)$ from both sides, dividing by δt and taking the limit as $\delta t \to 0$ gives

$$\frac{\mathrm{d}P_k(t)}{\mathrm{d}t} = \lambda[P_{k-1}(t)-P_k(t)],\ k = 1, 2, \ldots$$

$$\text{and} \qquad \frac{\mathrm{d}P_0(t)}{\mathrm{d}t} = -\lambda P_0(t). \tag{7.1}$$

The initial conditions are $P_0(0)=1$ and $P_k(0)=0$ for $k=1, 2, \ldots$ and the differential-difference equations (7.1) can be solved by generating function or other methods to give

$$P_k(t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

Thus, writing $\lambda t = m$, the mean number of vehicles expected in a given time, the Poisson distribution gives the probability of any given number, k, of vehicles passing during that time as

$$P(k) = \frac{e^{-m}.m^k}{k!}, \quad k = 0, 1, 2, \ldots. \tag{7.2}$$

If n equal periods of time are considered, the expected number of periods in which k vehicles arrive is given by

$$f(k) = \frac{n\,e^{-m}.m^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

This distribution fits in well with observation for light flows of, say, from 260 to 600 v.p.h. For example, for a flow of 300 v.p.h., the mean

TABLE 7.1

k	$P(k)$	$100\,P(k)$
0	0·006738	0·7
1	0·033690	3·4
2	0·084224	8·4
3	0·140374	14·0
4	0·175467	17·6
5	0·175467	17·6
6	0·146223	14·6
7	0·104445	10·5
8	0·065278	6·5
9	0·036266	3·6
10	0·018133	1·8
11	0·008242	0·8
12	0·003434	0·4
13	0·001320	0·1
14	0·000471	0·0
15	0·000228	0·0
Tota	1·000000	100·0

number expected in one minute is 300/60 = 5, so that the probabilities of 0, 1, 2, ... vehicles arriving in 1 min are given by

$$P(k) = \frac{e^{-5}.5^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

These probabilities are tabulated in Table 7.1: the entry 100 $P(k)$ gives the numbers of intervals out of 100 in which k vehicles arrive. A diagram of the corresponding discrete distribution is given in Figure 7.1. Note that the most probable numbers of arrivals are four and five.

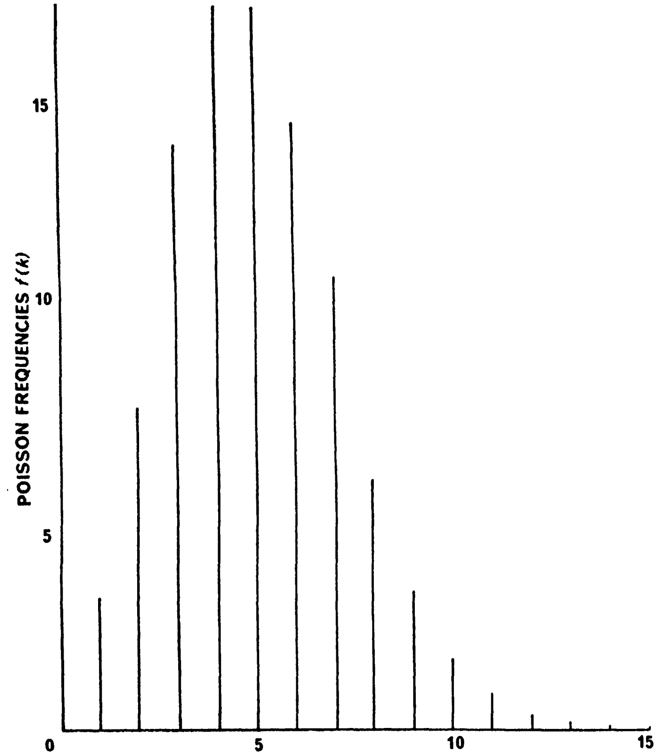

FIG. 7.1. Poisson frequencies $f(k) = \dfrac{100\,e^{-5}\,5^k}{k!}$

The other distribution – the negative exponential – gives the distribution of the lengths of intervals between vehicles, or headways, and is perhaps even more important than the Poisson in traffic flow. This distribution is now derived.

Take an event (arrival of a vehicle) as origin and let T be the time up to the next event with

$$\Phi(t) = P(T > t),$$

so that the cumulative distribution function (c.d.f.). is

$$1-\Phi(t)=P[T \leqslant t]=F(t)$$

Then $\Phi(t+\delta t) = P[\text{No event in } (0,t) \text{ and no event in } (t,t+\delta t)]$

$$= \Phi(t)[1-\lambda\delta t+o(\delta t)]$$

i.e. $\Phi'(t) = -\lambda\Phi(t)$

giving $\Phi(t) = \mathrm{e}^{-\lambda t}$, since $\Phi(0) = 1$.

Hence the probability density function (p.d.f.) of intervals is given by

$$f(t) = \mathrm{d}\,F(t) = -\Phi'(t) = \lambda\mathrm{e}^{-\lambda t},$$

where λ is the mean number of vehicles arriving in unit time, as in the Poisson distribution just derived.

We can now deduce a number of results. The probability of the time-interval or gap between two successive vehicles being less than t is

$$P[\text{gap} \leqslant t] = \int_0^t \lambda\mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1-\mathrm{e}^{-\lambda t}$$

and $$P[\text{gap} > t] = \int_t^\infty \lambda\mathrm{e}^{-\lambda t}\,\mathrm{d}t = \mathrm{e}^{-\lambda t}.$$

The sum of these probabilities is, of course, unity. Using these results we can plot the probability of a gap $\geqslant t$ sec against length of gap in seconds for various traffic volumes. For instance, for a flow of 600 v.p.h., $\lambda = 600/3{,}600 = 1/6$ for times measured in seconds, and we get the figures given in Table 7.2.

Consideration of this table shows that out of 100 time headways about half are greater than 4 sec and about a quarter greater than 8 sec. These figures have been plotted in Fig. 7.2. In practice, for vehicles travelling at 50 m.p.h. the average minimum spacing is 140 ft in daytime, i.e. less than 2 sec. Smeed and Bennet (1948) have

TABLE 7.2

t	$P(\text{gap} \geqslant t)$	$100P$	t	$P(\text{gap} \geqslant t)$	$100P$
0·5	0·9201	92	10·5	0·1738	17
1·0	0·8465	85	11·0	0·1599	16
1·5	0·7788	78	11·5	0·1471	15
2·0	0·7166	72	12·0	0·1353	14
2·5	0·6592	66	12·5	0·1245	12
3·0	0·6065	61	13·0	0·1146	11
3·5	0·5581	56	13·5	0·1054	10
4·0	0·5134	51	14·0	0·0970	10
4·5	0·4724	47	14·5	0·0892	9
5·0	0·4346	43	15·0	0·0821	8
5·5	0·3998	40	15·5	0·0755	8
6·0	0·3679	37	16·0	0·0695	7
6·5	0·3385	34	16·5	0·0639	6
7·0	0·3114	31	17·0	0·0588	6
7·5	0·2865	29	17·5	0·0541	5
8·0	0·2636	26	18·0	0·0498	5
8·5	0·2425	24	18·5	0·0458	5
9·0	0·2231	22	19·0	0·0421	4
9·5	0·2053	21	19·5	0·0388	4
10.0	0·1889	19	20·0	0·0357	4

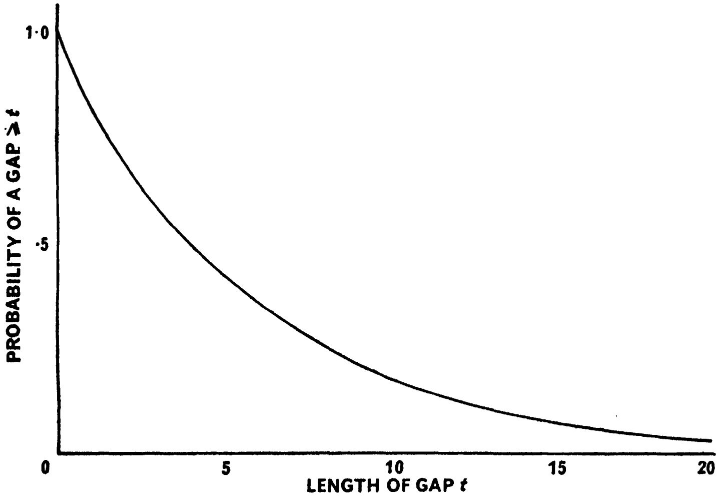

Fig. 7.2. Negative exponential probabilities $\Phi(t) = e^{-t/6}$

stated that a minimum time headway is 1·3 sec at 50 m.p.h., and that fairly constant headways of 1·3 to 2·0 sec are found for speeds from 0 to 60 m.p.h. Other data based on the reaction times of drivers give spacings of 40–1,230 ft for a range of speeds of (20–60) m.p.h. A 1-sec headway is therefore the minimum that it is sensible to consider.

Note that the mean time headway is $\bar{t} = 1/\lambda$ so that we can write

$$\Phi(t) = e^{-t/\bar{t}}.$$

If we wish to eliminate very small headways it is only necessary to use a shifted negative exponential distribution

$$\Psi(t) = e^{-(t-\tau)/(\bar{t}-\tau)}, \quad t \geqslant \tau = 1{\cdot}3 \text{ sec, say.}$$

Similar calculations have been carried out for various traffic volumes, viz. 100(100)1,000 v.p.h. and the results are summarized in Table 7.3.

TABLE 7.3

Flow v.p.h.	Probabilities gap $\geqslant t$		
	$t=4$	$t=8$	$t=10$
100	0·8948	0·8008	0·7574
200	0·8008	0·6412	0·5737
300	0·7166	0·5134	0·4346
400	0·6412	0·4111	0·3292
500	0·5737	0·3292	0·2493
600	0·5134	0·2636	0·1889
700	0·4594	0·2111	0·1431
800	0·4111	0·1690	0·1084
900	0·3679	0·1353	0·0821
1000	0·3292	0·1084	0·0622

We can see from this table that the probabilities of large headways decrease rapidly as the flow increases.

We conclude with a few simply derived theoretical results which are of interest in problems involving pedestrian delays and delays to minor-road vehicles at a T-junction.

(i) Mean time headway for all vehicles

$$= \int_0^\infty t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t$$

$$= [-t\mathrm{e}^{-\lambda t}]_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda t}\,\mathrm{d}t$$

$$= \frac{1}{\lambda}.$$

This is as one would expect.

(ii) Proportion of intervals $\leqslant t$ sec

$$= \int_0^t \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1 - \mathrm{e}^{-\lambda t}.$$

Proportion of intervals $> t$ sec

$$= \int_t^\infty \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \mathrm{e}^{-\lambda t}.$$

(iii) Proportion of time occupied by intervals $\leqslant t$ sec

$$= \frac{\int_0^t t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}{\int_0^\infty t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}$$

$$= \frac{[-t\mathrm{e}^{-\lambda t} - \mathrm{e}^{-\lambda t}/\lambda]_0^t}{[-t\mathrm{e}^{-\lambda t} - \mathrm{e}^{-\lambda t}/\lambda]_0^\infty}$$

$$= \lambda\left[\frac{1}{\lambda} - t\mathrm{e}^{-\lambda t} - \frac{\mathrm{e}^{-\lambda t}}{\lambda}\right]$$

$$= 1 - \mathrm{e}^{-\lambda t}[1 + \lambda t].$$

Proportion of time occupied by intervals $> t$ sec

$$= \frac{\int_t^\infty t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}{\int_0^\infty t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}$$

$$= \mathrm{e}^{-\lambda t}[1 + \lambda t].$$

(iv) Mean time headway for all intervals $\leqslant t$

$$= \frac{\int_0^t t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}{\int_0^t \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}$$

$$= \frac{1}{\lambda} - \frac{t\,\mathrm{e}^{-\lambda t}}{1-\mathrm{e}^{-\lambda t}}.$$

Mean time headway for all intervals $> t$

$$= \frac{\int_t^\infty t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}{\int_t^\infty \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t}$$

$$= \frac{t\,\mathrm{e}^{-\lambda t} + \mathrm{e}^{-\lambda t}/\lambda}{\mathrm{e}^{-\lambda t}}$$

$$= t + \frac{1}{\lambda}.$$

The Poisson or random hypothesis for traffic was the earliest and most tractable one. However, it has been shown by several authors that only when the rate of flow of vehicles is very light and vehicles can overtake freely, are they approximately randomly distributed on a road. Generally, overtaking is partially restricted either by the presence of other vehicles in the overtaking lane or by bends, brows of hills and so on which prevent drivers from seeing whether the road is clear for overtaking. Also there are usually some drivers who, upon catching up slightly slower vehicles, are content to follow behind. Hence there is almost always some bunching of vehicles. To some extent, therefore, the simple random hypothesis has been replaced by more sophisticated ones, for example the idea of the random queues of the Borel-Tanner (1961) or Miller (1961) distributions. In the latter the intervals between the heads and tails of queues or bunches of vehicles follow a negative exponential distribution. These distributions have more complicated mathematical forms than those described above.

For further information on the Borel-Tanner and Miller distributions the reader is referred to the relevant papers. We will give here only two formulae. The probability that a bunch of vehicles is of size n is given by Borel as

$$p_n = \frac{e^{-nr}(nr)^{n-1}}{n!}, \quad n = 1, 2, 3, \ldots.$$

where r is the traffic intensity. This quantity is equal to kd, where k is the concentration of vehicles or average number per unit length of road and d is the distance apart of the vehicles. The process can otherwise be thought of as a queue with random arrivals, rate k, and constant service time d. The moment generating function (m.g.f.) of this distribution has been given by Tanner (1953), and from this one can deduce the expected (or average) value of n. This is

$$E(n) = \frac{1}{(1-r)}.$$

Tanner extended this concept to the general case in which the queue originally consists of c units.

As well as being mathematically more tractable the random vehicles model has the advantage also that it requires knowledge of only one attribute of vehicular traffic, viz. the number of vehicles passing per unit time, an easy quantity to measure. The random queues model requires three attributes of the flow, the number of queues passing per unit time, the mean number of vehicles in the queue and the mean distance apart of vehicles in a queue. All these are difficult to measure and are different on different roads.

At many sites vehicle arrivals do closely approximate a series of random instants, but traffic in urban areas often departs from this rule for many reasons – for example, the presence of traffic lights.

Take, for instance, a cross-roads with traffic signals having a cycle-time of 60 sec. If the signals are fully saturated and have a 30-sec green phase, there will be a flow at a fixed level leaving the signal for approximately this period, and for the remaining 30 sec of red there will be a considerably smaller flow due to turning vehicles. The flow of vehicles away from a red signal is one quarter of that from a green at a typical cross-oroads in London. As the vehicles travel down the road they spread out in time: the rectangular waves are smoothed off until they eventually die away. In Oxford Street, London, the linked signals are spaced at about 500 ft, and at this distance arrivals are not random. At 1,000 ft the range of variation of the flow is reduced to 62% of its initial value, but there is still a marked 'wave' effect. At 1,500 ft the variation is 25% and is almost lost in the random fluctuations.

Some urban factors, on the other hand, make traffic more random. Vehicles leave queues from random positions, for example, and at random points on the road. They also join the stream at random points. The random vehicles model gives good results in many instances though it usually overestimates the number of small intervals. One solution that has been suggested is that a double distribution should be used, a normal distribution or a gamma distribution with small variance for small intervals (up to about 9 sec, say) and a negative exponential for large intervals.

In other cases where traffic is bunched and there is a preponderance of very short and very long headways the Borel distribution may be more realistic – but only under certain conditions. The roads must have fairly uniform characteristics, i.e. they must be of uniform width and either continuously straight or uniformly winding. If a road has both straight and winding sections, or narrow and wide sections, overtaking will tend to occur at certain sections of the road (at the wider or straighter sections) and downstream from these sections the queues themselves will occur in bunches.

7.2 Pearson's Distributions

Pearson originally derived his twelve distributions from the integration of a differential equation. The two most important of these nowadays are the Type I and the Type III distributions:

Type I $\quad p(x) = C(x-\alpha)^{\nu}(\beta-x)^{\mu}, \quad \alpha \leqslant x \leqslant \beta;$

Type III $\quad p(x) = C_1 x^{\nu-1}\,\mathrm{e}^{-\mu x}, \quad x \geqslant 0.$

The Type I distribution is perhaps better known now as the beta distribution. If we put $\alpha=0$, $\beta=1$, $\nu=p-1$, $\mu=q$ we get

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}, \quad 0 < x < 1, p > 0, q > 0.$$

This has mean $p/(p+q)$ and variance $pq/[(p+q)^2(p+q+1)]$ and the cumulative distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x f(x) = \dfrac{\beta(p,q,x)}{B(p,q)}, & 0 \leqslant x \leqslant 1 \\ 1, & x > 1. \end{cases}$$

where $\beta(p,q,x)$ is the incomplete beta function.

The Type III distribution is otherwise known as the Erlang or gamma distribution. Writing q for μ and p for ν we get

$$f(x) = \frac{q^p x^{p-1} \mathrm{e}^{-qx}}{\Gamma(p)}, \quad 0 < x < \infty, p > 0.$$

This has mean p/q and variance p/q^2. The cumulative distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{\gamma(p,q,x)}{\Gamma(p)}, & x \geqslant 0 \end{cases}$$

where $\gamma(p,q,x)$ is the incomplete gamma function.

Putting $p=1$ in this latter distribution gives the particular case of the negative exponential distribution with mean $1/q$ and variance $1/q^2$. The shape of the gamma distribution is influenced to a considerable extent by the choice of p and q, the two parameters, as shown in Fig. 7.3.

Many analyses have shown that distributions of velocity are slightly skew. If the Normal distribution will not do, therefore, a gamma distribution with the value of p of the order 25 may well give a good fit. This value of p is determined as follows. Under fairly free-flowing conditions the coefficient of variation of the speed distribution, i.e. the ratio of standard deviation to mean speed, is about 18%. Hence p is given by $1/\sqrt{p}=0{\cdot}18$. The distribution of the minimum gap acceptable to a driver wishing to cross a T-junction or a pedestrian wishing to cross a road (*vide* Chapter 8) may also be taken as a gamma distribution. This distribution represents a situation midway between randomness and regularity.

The generalized Poisson distribution has the same relation to the gamma distribution as the Poisson has to the negative exponential. This means that when the gaps between points on a line are distributed in a gamma distribution, the corresponding probability of n points in an interval of length x is the generalized Poisson probability with parameter qx.

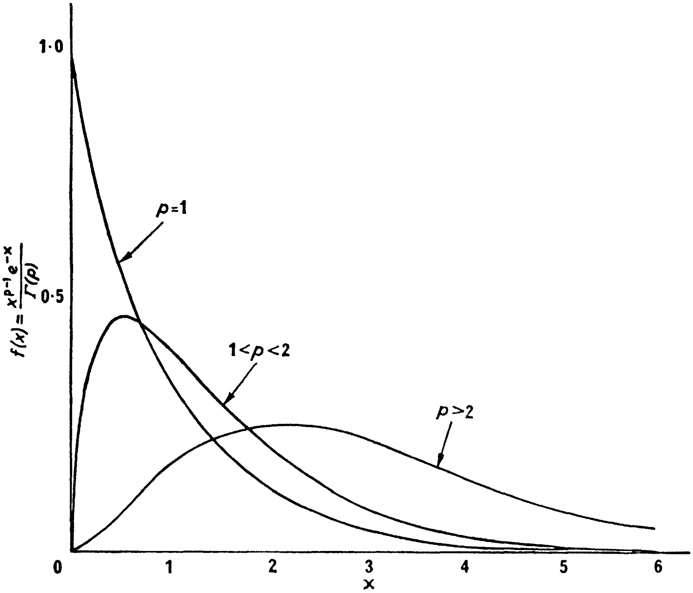

FIG. 7.3. Gamma distribution $f(x) = \dfrac{q^p\, x^{p-1}\, e^{-qx}}{\Gamma(p)}$ for various values of p.

The generalized Poisson distribution, parameter q, is given by

$$P_n = \sum_{i=1}^{p} \frac{e^{-q}\, q^{(np+i-1)}}{(np+i-1)!}, \quad n = 0, 1, 2, \ldots$$

where p is some positive integer. This means that the probability that $n=0$ is the sum of the first p terms of the Poisson series. The probability that $n=1$ is the sum of the next p terms and so on. The parameter q is equal to the mean only if $p=1$: this is, of course, the familiar case of the random distribution.

In the study of vehicular traffic the gamma distribution for gaps has been found to be rather more satisfactory than the random distribution in the description of traffic. The other is that there is always some interference between vehicles which induces an element of regularity into arrival times. With suitable adjustment of p and q the gamma distribution can be made to fit a wide variety of data. Further, this distribution has been used in another context (*vide* the Erlangian distribution of queueing theory) to describe a situation intermediate between regular and random arrivals.

7.3 Queueing Theory

Queueing theory is now more than 60 years old, having begun around 1900 with the work of Erlang on telephone traffic. It is still being extended, and in fact it is the most fully developed mathematical theory yet available for dealing with problems of congestion. The results of this theory have, however, only been applied to road traffic in very recent years because formulation in traffic terms is very difficult.

Vehicular traffic differs from the models associated with telephone traffic in at least three respects. Considered as a stochastic process the positions of vehicles in space and time are not identical: this is a consequence of variability in speeds. There is also the question of the finite size of vehicles: this gives rise to local interference and congestion due to interaction. Speed of travel is finite and vehicles do not move with complete freedom. They are subject to various controls: they must try to avoid collisions, for example. The problems of traffic networks are quite different from those of telephone networks. Optimal scheduling in the traffic context is meaningless. In addition the numbers of both vehicles and intersections is much greater than in classical network problems.

The first step is to provide adequate capacity for the average flow through the system, otherwise there is constant congestion. However, even if the capacity is adequate in this sense, congestion can still occur because the flow fluctuates. Queueing theory describes these fluctuations.

Congestion in this latter sense can arise in two ways. In one case there is a restriction on service in that not more than a limited number of customers can be served at one time, and congestion arises because the unserved customers must queue up and await their turn. This would be the case if a queue of vehicles were waiting at a toll-bridge, for example. In the other case there is the restriction that service is only available during limited periods. This would be the case if pedestrians were waiting to cross a road at a pedestrian crossing. Service is available only when a sufficiently large gap occurs in the traffic, but when it does, a large number of pedestrians can cross simultaneously. If the customers are vehicles waiting at a T-junction service is restricted in both senses, since a vehicle has to wait first whilst preceding vehicles cross and then for its own gap in the major road traffic.

The practical aim in investigating a system having congestion is, of course, to change the system in such a way as to ameliorate the situation. Queueing theory may be able to predict what amount of congestion is likely to occur due to a series of modifications, and thus help in choosing the best one, without the necessity for actual experiments arising.

The two main properties of a queueing system which are of practical importance are:

(i) The mean and distribution of the length of time a vehicle is in a queue, and of the total waiting time, which is equal to the queueing time plus the service time.

(ii) The mean and distribution of the number of vehicles in the system at any time.

Of these quantities the mean values are the ones most often needed.

One or two of the simplest results will be considered here: for a more extended treatment see, for example, Queues by Cox and Smith (1961).

There are three main characteristics of the system which must be specified:

(i) Arrival pattern–average arrival rate and distribution of gaps.

(ii) Service distribution–average service time and number of channels.

(iii) Queue discipline, e.g. first come, first served.

Let us consider the simple queue with random arrivals and departures. This means that the arrivals are distributed in a Poisson distribution with parameter λ, say. Since the departures are also random this implies that the service-time distribution will be negative exponential with mean $1/\mu$, say.

Let $p_n(t)$ denote the probability that the queue is in state n (i.e. consists of n vehicles) at time t and consider time $(t+\delta t)$. State n can be attained in the following three ways:

(i) State n, no arrival, no departure, and the probability of this is

$$p_n(t)[1-\lambda\delta t][1-\mu\delta t]+o(\delta t).$$

(ii) State $(n-1)$, one arrival, no departure, and the probability of this is

$$p_{n-1}(t)\,\lambda\delta t(1-\mu\delta t)+o(\delta t).$$

(iii) State $(n+1)$, one departure, no arrival, and the probability of this is

$$p_{n+1}(t)\,\mu\delta t(1-\lambda\delta t)+o(\delta t).$$

Then, summing over all the probabilities

$$\begin{aligned} p_n(t+\delta t) &= p_n(t)(1-\lambda\delta t)(1-\mu\delta t)+p_{n-1}(t)\,\lambda\delta t(1-\mu\delta t)+ \\ &\qquad +p_{n+1}(t)\,\mu\delta t(1-\lambda\delta t)+o(\delta t), \quad n>0 \end{aligned}$$

and

$$p_0(t+\delta t) = p_0(t)(1-\lambda\delta t)+p_1(t)\,\mu\delta t(1-\lambda\delta t)+o(\delta t)$$

giving

$$\begin{aligned} \frac{p_n(t+\delta t)-p_n(t)}{\delta t} &= -(\lambda+\mu)\,p_n(t)+\lambda p_{n-1}(t)+\mu p_{n+1}(t) \\ &\qquad +o(\delta t)/\delta t, \quad n>0 \end{aligned}$$

and

$$\frac{p_0(t+\delta t)-p_0(t)}{\delta t} = -\lambda p_0(t)+\mu p_1(t)+o(\delta t)/\delta t.$$

As $\delta t\rightarrow 0$ these reduce to the set of differential-difference equations

$$\begin{aligned} \frac{\partial}{\partial t}p_0(t) &= -(\lambda+\mu)\,p_n(t)+\lambda p_{n-1}(t)+\mu p_{n+1}(t), \quad n>0 \\ \frac{\partial}{\partial t}p_n(t) &= -\lambda p_0(t)+\mu p\ (t). \end{aligned}$$

Using the generating function method of solution, put

$$\pi(s,t) = \sum_{n=0}^{\infty} p_n(t)\,s^n, \quad 0<s<1. \tag{7.3}$$

Multiplying the nth equation of the set by s^n and adding gives

$$\frac{\partial}{\partial t}\pi(s,t) = \pi(s,t)\left[\mu\frac{1-s}{s}-\lambda(1-s)\right]-\mu\frac{1-s}{s}p_0(t).$$

This equation can be solved quite generally, but we are mainly interested in the case where the queueing process has settled down: this is the equilibrium state. For the equilibrium solution we make the assumption that the probabilities are not time-dependent, so that the left-hand side equals zero and $\pi(s,t)$ becomes $\pi(s)$.

Then
$$\pi(s)(\mu-\lambda s)-\mu p_0 = 0$$

i.e.
$$\pi(s) = \frac{p_0}{1-s\rho}, \tag{7.4}$$

where $\rho=\lambda/\mu$ is the traffic intensity.
Putting $s=1$ in equation (7.3) gives

$$\pi(1) = \sum_0^\infty p_n(t) = 1,$$

and substitution in equation (7.4) then gives

$$p_0 = 1-\rho$$

and hence
$$\pi(s) = \frac{1-\rho}{1-s\rho}.$$

Now $\pi(s)$ is the generating function of the geometric distribution and p_n is the coefficient of s^n in the expansion of $\pi(s)$, viz.

$$p_n = \rho^n(1-\rho), \quad n = 0, 1, 2, \ldots.$$

These are the asymptotic (or long-term) queue-length probabilities, and the mean queue-length is given by

$$(1-\rho)\sum_1^\infty n\rho^n = (1-\rho)(\rho+2\rho^2+3\rho^3+\ldots)$$

$$= \frac{\rho}{1-\rho}.$$

Similarly the variance is given by

$$(1-\rho)\sum_1^\infty n^2\rho^n-\frac{\rho^2}{(1-\rho)^2} = \frac{\rho(1-\rho)(1+\rho)}{(1-\rho)^3}-\frac{\rho^2}{(1-\rho)^2}$$

$$= \frac{\rho}{(1-\rho)^2}.$$

If $\rho < 1$, $\pi(s)$ is a probability generating function with $p_0 = 1 - \rho$, an equilibrium solution exists and the queue is stable. If $\rho \geqslant 1$, $\pi(s)$ is not a probability generating function since the terms are either zero or negative. If $\rho > 1$ no equilibrium solution exists, the queue is unstable and grows indefinitely. Qualitatively the case $\rho = 1$ is more difficult to describe, but for practical purposes should be classed with $\rho > 1$. For ρ near unity minor effects such as slight dependence of arrival rate on queue size may decide whether or not an equilibrium state exists or not.

If $\rho < 1$ the variance is greater than the mean, and as $\rho \to 1$, i.e. as the queue approaches saturation, the variance becomes very large. In practical terms this means that the average queue-length is subject to enormous fluctuations which can lead to substantial waiting times. This has been confirmed by simulation studies, but for further discussion of this phenomenon see Chapter 9.

The probability of finding more than n vehicles in the queue is ρ^{n+1}, hence if ρ is small long queues are very unlikely to occur. Table 7.4 gives the mean queue sizes and the probabilities of more than five vehicles in the queue for various traffic intensities.

TABLE 7.4

ρ	Mean queue-length	Probability that queue-length exceeds 5
0·1	0·111	0·000
0·2	0·250	0·000
0·3	0·429	0·001
0·4	0·667	0·004
0·5	1·000	0·016
0·6	1·500	0·047
0·7	2·333	0·118
0·8	4·000	0·262
0·9	9·000	0·531

Let us now consider the waiting time of an individual vehicle, defined so as to exclude its own service-time. Consider a queue at time t and let $w(x,t)$ be the probability density function of waiting time so that the cumulative distribution function is $W(x,t)$ where

$$W(x,t) = \int_0^x w(x,t)\,\mathrm{d}x = \text{Prob (waiting time} \leqslant x \text{ at } t).$$

Now the queue is empty with probability $(1-\rho)$, and in this case the waiting time is zero. Hence $w(x,t)$ has a discrete component at the origin of height $(1-\rho)$. When the queue is not empty the waiting time is equal to the sum of the service-times of the $(n-1)$ vehicles waiting plus the residual service-time of the vehicle in service. Now if the distribution of service-times is negative exponential the waiting time distribution is the sum of $(n-1)$ negative exponentials plus the distribution of residual waiting time.

The latter distribution is also negative exponential since it is an inherent property of randomness that the distribution of the length of gap is the same whether or not a length has been removed. This can easily be shown in the following way:

If Prob $(X \leqslant x) = 1-\mathrm{e}^{-\mu x}$, then the conditional probability is given by

$$\begin{aligned}\text{Prob }(X \leqslant x|X > x_0) &= \frac{\text{Prob }(x_0 < X < x)}{\text{Prob }(X > x_0)} \\ &= \mathrm{e}^{\mu x_0}\int_{x_0}^{x} \mu\,\mathrm{e}^{-\mu z}\,\mathrm{d}z \\ &= 1-\mathrm{e}^{-\mu(x-x_0)},\end{aligned}$$

and with x_0 taken for the new time origin this is $1-\mathrm{e}^{-\mu x}$.

Hence the conditional distribution of waiting time $w(x|n)$, is the sum of n negative exponential distributions. There are a number of ways of obtaining the distribution of a sum such as this: we shall use moment generating function methods here. Now the moment generating function for a negative exponential with parameter μ is given by

$$\begin{aligned}M(t) = E(\mathrm{e}^{tx}) &= \int_0^\infty \mathrm{e}^{tx}\,\mu\,\mathrm{e}^{-\mu x}\,\mathrm{d}x \\ &= \frac{\mu}{\mu-t} \\ &= \left(1-\frac{t}{\mu}\right)^{-1}.\end{aligned}$$

A well-known theorem states that the m.g.f. of the sum of n independent variates is the product of the n individual m.g.f.'s. Hence in this case the m.g.f. for the distribution of the sum of n negative exponential variates is

$$\left(1-\frac{t}{\mu}\right)^{-n},$$

and this can easily be shown to be the m.g.f. of a gamma distribution.

Hence $$w(x|n) = \frac{\mu^n}{\Gamma(n)} e^{-\mu x} x^{n-1}.$$

However, the probability of a queue having n members is $(1-\rho)\rho^n$ in equilibrium. Hence the unconditional distribution of waiting-time, adding in the discrete component, is given by

$$w(x) = (1-\rho)\,\delta(x) + \sum_{n=1}^{\infty} \frac{\mu^n e^{-\mu x} x^{n-1}}{\Gamma(n)} (1-\rho)\,\rho^n$$

where $\delta(x)$ is the Dirac delta function. Then since $\rho = \lambda/\mu$,

$$\begin{aligned} w(x) &= (1-\rho)\delta(x) + e^{-\mu x}(1-\rho)\,\lambda \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} \\ &= (1-\rho)\,\delta(x) + \rho(\mu-\lambda)\, e^{-\mu x+\lambda x} \qquad (7.5) \end{aligned}$$

or, integrating,

$$W(x) = 1-\rho\, e^{-\mu(1-\rho)x}.$$

We then have, multiplying equation (7.5) by x and integrating from 0 to ∞,

$$\begin{aligned} \text{mean } w(x) &= \int_0^{\infty} (1-\rho)\, x\,\delta(x)\, dx + \rho(\mu-\lambda) \int_0^{\infty} x\, e^{-(\mu-\lambda)x}\, dx \\ &= \rho(\mu-\lambda)\left\{\left[\frac{x\, e^{-(\mu-\lambda)x}}{\lambda-\mu}\right]_0^{\infty} - \frac{1}{(\lambda-\mu)} \int_0^{\infty} s^{-(\mu-\lambda)x}\, dx\right\} \\ &= \frac{\rho}{\mu(1-\rho)}. \end{aligned}$$

Similarly

$$\begin{aligned}\operatorname{var} w(x) &= \int_0^\infty w(x)\,x^2\,\mathrm{d}x-\frac{\rho^2}{\mu^2(1-\rho)^2}\\ &= \int_0^\infty (1-\rho)\,x^2\,\delta(x)\,\mathrm{d}x+\rho(\mu-\lambda)\int_0^\infty x^2\,\mathrm{e}^{-(\mu-\lambda)x}\,\mathrm{d}x\\ &\qquad -\frac{\rho^2}{\mu^2(1-\rho)^2}\\ &= \frac{2\rho}{(\mu-\lambda)^2}-\frac{\rho^2}{\mu^2(1-\rho)^2}\\ &= \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}.\end{aligned}$$

More general cases in which, for example, the service-time is no longer distributed in a negative exponential, are dealt with in the reference already given. The results are derived by considering the queueing system only at the moments of departure of a vehicle. It can be shown that if q_n is the number of vehicles in the queue at the time the service of the nth vehicle terminates, then the mean number of vehicles in the queue (assuming random arrivals) is given by

$$E(q_n) = \rho+\frac{\rho^2(1+C^2)}{2(1-\rho)},$$

where C is the coefficient of variation of the service time distribution, i.e. the ratio of its standard deviation to its mean. If the service is exponential, $C^2=1$ and then

$$E(q_n) = \frac{\rho}{1-\rho}. \tag{7.6}$$

If the service is constant, $C^2=0$ and then

$$E(q_n) = \frac{\rho(1-\frac{1}{2}\rho)}{(1-\rho)}. \tag{7.7}$$

It can also be shown that the mean waiting time in the latter case is

$$\frac{\rho}{2\mu(1-\rho)},$$

which is half of the mean queueing time for random service.

The ratio of the expressions in equations (7.7) and (7.6) is $1-\frac{1}{2}\rho$. Hence if ρ is small there is very little reduction in mean queue size if a change is made from random to regular service, but the reduction is about 50% as ρ approaches unity.

Notice the following interesting comparison, which is rather difficult to reconcile intuitively. We found previously, in the simple analysis for random arrivals and departures, that the mean queue size for the equilibrium state was $\rho/(1-\rho)$. This result referred to an arbitrary moment in time. We have found the same result for the mean queue size in the case where the time instants considered are those at which a vehicle has just completed service. It can also be shown that the distribution of queue size at moments when a service has just been completed is the same geometric distribution as the one that applied to queue size at an arbitrary point. In fact this turns out to be a property of queues having random arrivals only, and it is not found in more general cases.

The foregoing results are summarized in Table 7.5. If we wish to use these results in practice for the purpose of predicting queue-lengths and waiting-times at T-junctions, for example, we need to know the mean arrival rate of vehicles on the minor road and the mean service-time of these vehicles, which is of course dependent on the arrival rate on the major road. These calculations are usually carried out as follows. The intervals between arrivals are written down and the successive cumulative totals calculated, and from these the mean values, as follows

TABLE 7.5

Service	Mean queue-length	Mean waiting time
Random	$\rho/[(1-\rho)]$	$\rho/[\mu(1-\rho)]$
Regular	$[\rho(2-\rho)]/[2(1-\rho)]$	$\rho/[2\mu(1-\rho)]$

Intervals between arrivals: 4, 5, 10, 3, 5, 2, 8, 6, 4, ...
Cumulative totals: 9, 19, 22, 27, 29, 37, 43, 47, ...

Mean values $\frac{9}{2}, \frac{19}{3}, \frac{22}{4}, \frac{27}{5}, \frac{29}{6}, \ldots$

i.e. 4·5, 6·3, 5·5, 5·4, 4·9, ...

If the means are plotted against the number of intervals we obtain a broken line which settles down to some value which can be taken as the mean arrival rate, and similarly for service-times.

Table 7.6 gives the mean queue-lengths for random and regular service for different values of the traffic intensity. Tables 7.5 and 7.6 show that, for both queue-lengths and waiting times, regular service is better for the customer. For $\rho > 0{\cdot}8$ the mean waiting times increase rapidly with any increase in ρ, hence a slight increase in demand where traffic intensity is already high can lead to a bottleneck.

TABLE 7.6

Traffic intensity ρ	Mean queue-lengths Random $\rho/(1-\rho)$	 Regular $\rho(2-\rho)/2(1-\rho)$
0·1	0·111	0·106
0·2	0·250	0·225
0·3	0·429	0·364
0·4	0·667	0·533
0·5	1·000	0·750
0·6	1·500	1·050
0·7	2·333	1·517
0·8	4·000	2·400
0·9	9·000	4·950

The sort of theoretical analysis reviewed here could be applied, for instance, to the situation of traffic queueing at a toll booth at the entrance to a toll-bridge. Queues may form because of the length of time taken to pay the toll (service-time) and because of variation in arrival times. We could measure the mean rate of flow over the bridge (μ) and hence find the mean time spent in paying toll ($1/\mu$) and the probability distribution of service-time. Similarly we could measure the mean arrival rate (λ) and the probability distribution of time intervals between arrivals. From this data we should be able to set up a mathematical model of the queue. Three cases are of interest:

(i) Service and arrivals both regular. This is a trivial case and there are two alternatives:

$\lambda \leqslant \mu$: no queue;

$\lambda > \mu$: queue-length becomes infinite.

(ii) Both service and arrivals random: this is more realistic.

$\lambda \leqslant \mu: p_n = \rho^n(1-\rho)$;

mean queue-length $= \rho/\mu(1-\rho)$;

mean queueing + paying time $= 1/\mu(1-\rho)$

$\lambda > \mu$: queue becomes infinite and none of the formulae apply.

In practice if μ is less than 50% greater than λ, i.e. if $\rho < 0{\cdot}7$, the chance of encountering long queues is not great and the mean delay is twice the mean time to pay. When $\rho > 0{\cdot}8$ saturation effects set in and both the mean queue-length and the mean waiting time become infinite.

(iii) Service regular, arrivals random. The mean queue-length reduces to half that of (ii) but saturation effects are still present.

Let us now consider a matter which is of great significance in traffic engineering, namely that of the peak periods when traffic demand exceeds capacity. If $\rho < 1$ the queueing problem is solved by considering equilibrium solutions which are independent of time. For $\rho > 1$ it is necessary to consider the transient state in which there is a continuing build-up of traffic. In this case the mean queue-length, $E[q(t)]$, becomes infinite as t increases.

Suppose that, prior to the peak period, the system is in equilibrium with an initial traffic intensity $\rho_0 = \lambda_0/\mu$, and that the expected number of vehicles in the system is $E(q)$. Now suppose that ρ_0 increases to $\rho_1 = \lambda/\mu > 1$, in such a way that the service-time or departure rate μ is constant before and after the peak.

Consider the case where both arrivals and departures are random with rates λ and μ respectively. The mean number in the system at a time t after the peak starts is given by the equilibrium value increased by the number of arrivals and decreased by the number of departures in time t, thus

i.e.
$$E[q(t)] \approx E(q) + (\lambda - \mu)t$$
$$= E(q) + \mu(\rho_1 - 1)t. \tag{7.8}$$

Similarly the variance, summing over the three components, is given by

$$\operatorname{var}[q(t)] \approx \operatorname{var}(q) + \operatorname{var}[\lambda(t)] + \operatorname{var}[\mu(t)].$$

But $E(q)$ and $\text{var}(q)$ are given by the equilibrium analysis as

$$E(q) = \frac{\rho_0}{1-\rho_0}, \quad \text{var}(q) = \frac{\rho_0}{(1-\rho_0)^2}.$$

Also $\text{var}[\lambda(t)]$ and $\text{var}[\mu(t)]$ are the variances of the numbers arriving and departing (or being served) in time t, i.e. of λt and μt. Since the distributions are Poisson the variances are equal to the means, hence

$$E[q(t)] \approx \frac{\rho_0}{1-\rho_0} + \mu(\rho_1 - 1)t,$$

$$\text{var}[q(t)] \approx \frac{\rho_0}{(1-\rho_0)^2} + \mu(\rho_1 + 1)t.$$

If the service-times are constant instead of negative exponential equation (7.8) holds but with $E(q)$ now given by

$$E(q) = \frac{\rho_0(1-\frac{1}{2}\rho_0)}{1-\rho_0}.$$

Hence in this case

$$E[q(t)] = \frac{\rho_0(1-\frac{1}{2}\rho_0)}{(1-\rho_0)} + \mu(\rho_1 - 1)t.$$

By way of an example consider a simple queue with a random arrival rate of 1 vehicle/min and a mean service-time of 40 sec. Then

$$\rho_0 = \frac{40}{60} = \frac{2}{3}.$$

Now let the arrival rate double to 2 vehicles/min so that

$$\rho_1 = \frac{4}{3},$$

and suppose that this rate is maintained during the succeeding hour, the service-time being unaltered and the arrivals still random. Then

$$E[q(60)] = \frac{2/3}{1-2/3} + \frac{3}{2}\left(\frac{4}{3} - 1\right)60 = 32$$

i.e. the expected value of the queue-length after 1h is 32.

In the case of a constant service-time of 40 sec, the corresponding formula gives

$$E[q(60)] = \frac{2(1-1/3)}{3(1-2/3)} + \frac{3}{2}\left(\frac{4}{3} - 1\right)60 = 31\tfrac{1}{3}.$$

In general it can be shown that the actual service-time distribution during peak hours is not critical. It should be noted that var $[q(t)]$ is very large (in this case ≈ 211) and the queue-length can oscillate wildly under these conditions: the average value, therefore, is not very meaningful.

The question now arises as to how long the peak-hour queue will take to dissipate when the traffic intensity has once more returned to its pre-peak value ρ_0.

Cox and Smith, in the reference already mentioned, make the assumption that service-time is constant, and that at the time when ρ returns to $\rho_0 < 1$ there are a large number of vehicles in the queue. They also define the point at which the queue is dissipated as that at which the queueing time of a newly-arrived vehicle is exactly equal to the average queueing-time of the vehicles when the system is in statistical equilibrium.

If T is the time to the dissipation of the queue from the point at which ρ becomes ρ_0, the equations developed by Cox and Smith are

$$E(T) = \frac{1}{(1-\rho_0)}\left[\frac{E[q(t)]}{\mu} - \frac{\rho_0}{2(1-\rho_0)}\right],$$

$$\operatorname{var}(T) = \frac{\rho_0}{\mu(1-\rho_0)^3}\left[\frac{E[q(t)]}{\mu} - \frac{\rho_0}{2(1-\rho_0)}\right] + \frac{1}{\mu^2}\left[\frac{\operatorname{var}[q(t)]}{(1-\rho_0)^2}\right].$$

Using the numerical values of the example above, these give

$$E(T) = 61 \text{ min},$$

$$\operatorname{var}(T) = 1111, \text{ whence } \sigma(T) \approx 33 \text{ min}.$$

Thus the effects of the rush-hour last for a further hour on average, with a standard deviation of about half an hour.

7.4 Application of Queueing Theory to Overtaking

Up to this point this chapter has dealt with the application of standard queueing theory to traffic situations. Before leaving the topic, mention should be made of the work of Miller (1961a, 1961b, 1962) on an overtaking model in which vehicles travel in random bunches or queues, and catch up and overtake other bunches at random times. In this model the velocities of vehicles follow a general probability distribution. Instantaneous changes of velocity are assumed and the vehicles are considered as points, i.e. there are no finite gaps between queueing vehicles. This is not a very realistic model as it stands, but some suggestions for corrections are included.

On the assumption that the rate of catching-up of queues is proportional to the product of the concentrations and their relative velocities, Miller derived a set of stochastic equations. He considered mainly stationary states in which the rate of catching-up was equal on average to the rate of overtaking. The slower queue in this model can be regarded as the service-point of classical queueing theory.

If k is defined as the concentration of queues of vehicles, and $k_1 < k$ that part of k made up of single vehicles, an approximate equation for the rate of change of concentration with respect to t can be derived. The rate at which queues catch one another up is proportional to the square of the concentration in this case, hence if ω is the rate of overtaking from queues consisting of more than one vehicle,

$$\frac{\mathrm{d}k}{\mathrm{d}t} = -ak^2 + \omega(k-k_1).$$

The constant a turns out to be half the mean difference in velocities, taken positive.

Assuming that the catching-up and overtaking has reached statistical equilibrium

$$\frac{k_1}{k} = 1 - \frac{ak}{\omega}.$$

If $k_1/k \sim 1$ nearly all the vehicles are travelling freely, and hence ak/ω can be used as a measure of bunching. If ω is large, there is little bunching unless k also is large. If ω is smaller there is bunching at lower concentrations.

Using this model Miller was able to study a number of problems. By considering queueing behind a slow vehicle he was able to estimate the interference caused to the traffic flow as a whole. He also applied the theory to queueing at fixed-cycle traffic signals and to the problem of waiting-time distributions for pedestrians wishing to cross a road.

The question of overtaking rates is intimately connected with that of road capacity. It is shown in a paper by Rallis (1965) that the values of 'basic' and 'practical' capacity for different types of roads, as given in the Highway Capacity Manual (Bureau of Public Roads, Washington, 1950), are directly related to percentage rates of failure to overtake or maintain free speed. For example, for a rural highway having two lanes in one direction the figures for basic and practical capacities correspond to a 40% and 22% failure to overtake.

This probability of failure to overtake, which is due to all the traffic lanes being occupied, can be calculated from a formula due to Erlang, which was derived originally for telephone traffic. This so-called 'loss' or 'rejection' formula in fact understates the degree to which traffic is affected by congestion. The probability that a vehicle suffers delay is greater than the probability that it fails to overtake, because arrivals accumulate and cause a prolongation of the condition. This probability is, however, more difficult to estimate because a number of assumptions are needed before actual numerical results can be obtained. The 'rejection' formula is not dependent on the distribution of traffic between the different lanes. However, an assumption of stationarity has to be made, and this involves an approximation to the practical situation.

If A is the traffic load and n the number of lanes, Erlang's formula, assuming Poisson arrivals, gives

$$R_n = \frac{A^n/n!}{\sum_{r=0}^{n} A^r/r!}.$$

This formula can be used to calculate any one of n, A and R_n if the other two are given. For example, if $n=2$ the corresponding values of R_2 for $A=0{\cdot}6$, $1{\cdot}0$ and $1{\cdot}2$ are 10%, 20% and 25%.

The traffic load $A=NT$, where N is the average number of vehicles/unit time and T is the time spent by the vehicle within the distance h, the safety distance between two vehicles. The speed $v=h/T$ and the concentration $k=N.T/h=A/h$.

The quantity R_n can be used in setting design standards. It can be used, for example, to determine whether a road needs the addition of an extra lane or not. The addition of one lane would reduce the number of road users suffering interference from NR_n/unit time to NR_{n+1}/unit time. In order to take into account economic considerations $N(R_n-R_{n+1})$ could then be equated to some function of the capital investment needed. This is Moe's principle.

Suppose, by way of an example, that it is desired to cater for a total traffic volume of 600 v.p.h. on a road having one lane in each direction; the mean speed of the traffic to be 30 m.p.h. and the minimum spacing 264 ft. We wish to find the percentage of vehicles unable to overtake and the improvement which can be obtained by adding an extra lane.

In this case
$$N = 600,$$
$$v = 30,$$
$$h = 1/20 \text{ ml},$$
$$n = 2.$$

Hence
$$A = \frac{600}{30} \times \frac{1}{20} = 1{\cdot}0,$$
$$R_2 = \frac{A^2/2}{1+A+A^2/2} = \frac{1/2}{5/2}$$
$$= 20\% \text{ or } 120 \text{ v.p.h.},$$
$$R_3 = \frac{A^3/6}{1+A+A^2/2+A^3/6} = \frac{1/6}{8/3}$$
$$= 5\tfrac{1}{2}\% \text{ or } 38 \text{ v.p.h.}$$

This means an improvement of $14\frac{1}{2}\%$ or a reduction of 62 v.p.h. in vehicles unable to overtake.

This sort of approach would enable design work to be placed on a rational basis.

CHAPTER 8

The Stochastic Approach to Static Delay Problems

8.1 Introduction

There has been a great deal of interest during the last 10 years in the problem of 'static delay'. This term is used here to describe the delay incurred at a stop sign or traffic signals by a vehicle wishing to cross, or turn into, a major-road at an intersection. An intersection having a stop sign is usually termed a priority intersection, whilst one controlled by traffic lights is a signal-controlled intersection. The problem of priority intersections is dealt with in §8.2 and §8.3 whilst that of signal controlled intersections is considered in §8.4. Problems involving delay on an open road are considered briefly in §8.5, and a new approach to the study of traffic behaviour at a bottleneck is described in §8.6.

Intersections are very important features of the road system, since by far the greatest loss attributable to a single cause occurs at these points. Their design has been the subject of a good deal of research. One of the most important questions to be answered by the traffic engineer relates to the kind of control to be used. Detailed studies have been carried out into the effect of different kinds of control on various factors, including delay, capacity and accident rates. Excessive delays to drivers are undesirable on economic grounds. Further, they can lead to acts of impatience and hence to accidents. In some circumstances however, the absorption capacity of the major-road is of more importance than the delay to minor-road vehicles, at any rate from the point of view of the traffic engineer. One example of this is the case of an exit from a theatre or cinema car-park.

Consider first the problem of priority intersections. In models of this problem it is assumed that the driver in the minor-road stops at the sign, estimates each inter-vehicle gap γ (measured in units of time), and makes a decision to cross or not to cross according as the gap is greater or less than some predetermined gap.

The problem of a single vehicle at a stop sign has already been fairly well documented, and the theory agrees reasonably well with observed behaviour. The problem is essentially the same as that of a group of pedestrians crossing a road, since such a group can cross together when an acceptable gap occurs. A more difficult problem is that of a queue of vehicles waiting at a stop sign. The simplest case refers to the problem of crossing a single lane of major-road traffic: the problem of crossing two lanes, one in each direction, is difficult to deal with mathematically. It should be noted also that in order that queueing theory should be applicable, the queue on the minor road is to be regarded as a single line of vehicles in which the original order of arrival is preserved.

There is a further complication in that many models of decision criteria for gap acceptance are possible. In some of the earlier work a step function gap acceptance was used, but in later work the gap γ has been taken as a random variable. Reliable experimental data is scarce; but personal experience suggests that the latter type of gap acceptance criterion is more realistic. Thus in this model, with each inter-vehicle gap γ is associated a probability $p(\gamma)$ that the driver will accept it. The probability $p(\gamma)$ increases monotonically with γ. Some observations have suggested considerable variability in the time-gap acceptable to drivers, but about 8 sec seems to be fairly typical.

In most work on this problem a single gap between two successive vehicles on the major-road is considered, and it is assumed that there are $n > 1$ vehicles waiting in the queue at the stop sign. The first driver considers the gap γ and makes a decision to cross on the basis of the gap acceptance criterion $p(\gamma)$. If the first driver decides to cross the road, the second driver moves up to the place vacated by the first and, assuming that no vehicle has yet passed on the major road, makes his decision on the basis of the gap acceptance function $p^*(\gamma - \gamma')$, where $p^*(x)$ may or may not be the same as $p(x)$, and where γ' is a time which depends on the time taken by the first driver to cross and also on the time taken by the second driver to move up one place to the front of the queue.

A good deal of empirical work would need to be done to determine the actual form of the function $p(\gamma)$. The result of such work would, no doubt, be interesting. However, it has been shown that the results of an analysis are not critically dependent on the exact form of the gap acceptance function. Three possibilities for this function are as follows:

(i) Step function, viz.

$$p(\gamma) = \begin{matrix} 0, & \gamma \leqslant \Gamma \\ 1, & \gamma > \Gamma. \end{matrix}$$

(ii) Trapezoidal function, viz.

$$p(\gamma) = \begin{matrix} 0, & \gamma \leqslant \gamma_0 \\ (\gamma-\gamma_0)/(\gamma_1-\gamma_0), & \gamma_0 < \gamma \leqslant \gamma_1 \\ 1, & \gamma > \gamma_1. \end{matrix}$$

(iii) Delayed exponential function, viz.

$$p(\gamma) = \begin{matrix} 0, & \gamma \leqslant \gamma_0, \\ 1-\exp[-\lambda(\gamma-\gamma_0)], & \gamma > \gamma_0 \end{matrix}$$

where $\lambda = 1/(\bar{\gamma}-\gamma_0)$ and $\bar{\gamma}$ = average gap accepted.

These are shown in Fig. 8.1.

Various authors have examined a number of special cases of the general queueing model and the associated problem of the distribution of gaps in traffic streams. Adams (1936) obtained a formula for the average wait for an unblocked interval in a random flow of traffic: this was a pedestrian crossing problem. Oliver (1962) studied the distribution of gaps in traffic streams with general probability distributions. Beckman, McGuire and Winsten (1956) have discussed a model in discrete time and obtained expressions for the moments of queue-lengths and waiting time. Oliver and Bisbee (1962) studied a model in which only one vehicle could be accepted at a given gap: this model corresponds to high density traffic on the major road. The minor road flow was random, but the major road distribution was arbitrary.

A more complete analysis of the queueing problem was given by Tanner (1962). He assumed that the flows were random on both roads, and the gap acceptance function was taken as a step function. The probability distribution of wait for an unblocked period was derived by him in three different ways. Weiss and Maradudin (1962) adopted a different approach: they used renewal theory. Jewell (1961, 1962) discussed various aspects of the queueing problem. His model was rather unrealistic in that a vehicle arriving at a stop sign when no other vehicles were present had to wait for a vehicle to pass on the major road before it could cross. Hawkes (1965) considered a model in which vehicles arrived in bunches at a junction but crossed one at a time. A signal showed red and green so as to form an alternating renewal process. He showed that this system could be generated

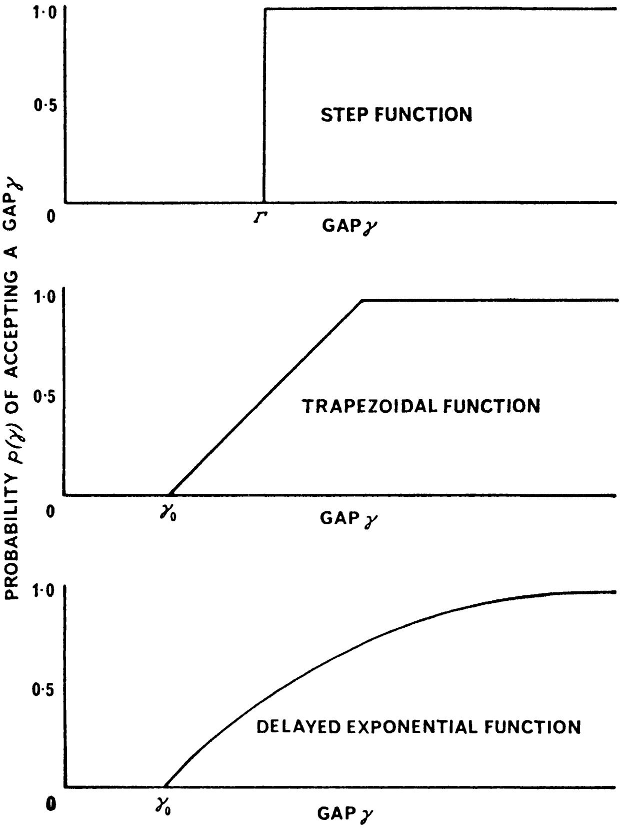

FIG. 8.1. Probability $p(\gamma)$ of accepting a gap γ.

by the major road traffic, and derived the equilibrium distribution of waiting time. Mention should also be made here of a book by Haight (1963) entitled 'Mathematical Models for Traffic Flow'. This is largely concerned with the stochastic approach to traffic flow problems.

The criterion for measuring the performance at an intersection is usually taken as the average delay to vehicles, but there are a number of other possible criteria. Some examples are the total vehicle-seconds delay over all the vehicles using the intersection, percentage of vehicles delayed by the system, average vehicle-seconds delay to delayed vehicles or average travel-time through the system. The problem arises, therefore, of what to optimize.

8.2 Simple Models of Priority Intersections

In Britain 80% of all fatal and serious accidents to pedestrians are connected with the act of crossing the road. To some extent the level of risk can be measured by the delay which a pedestrian experiences before he can cross. The delay is occasioned by his having to wait for a gap in the traffic, and these gaps occur in statistical distributions.

Suppose that a stream of traffic passes a point on the road in some probability distribution. To a pedestrian trying to cross the road, or to a driver on a minor road, the stream may be regarded as an alternating succession of periods which permit crossing and periods which do not. The probability distributions connected with these periods are developed and discussed below.

Consider first an analysis of a simple delay problem using the Laplace Transformation. Since pedestrians can cross as a bunch, a single vehicle and a bunch of pedestrians can, for the purpose of this analysis, be considered as the same thing. However, the analysis will be couched in terms of pedestrian delays.

The following notation is used in the remainder of this chapter. A probability density function will be denoted by p.d.f. and the corresponding cumulative distribution function by c.d.f. Let $f(t)$ denote a p.d.f. and $F(t)$ the corresponding c.d.f., $0 \leqslant t < \infty$ then the Laplace Transform of $f(t)$ is defined by

$$\mathscr{L}[f(t)] = \phi(s) = \int_0^\infty e^{-st} f(t)\,dt.$$

If $f(x)$, $F(x)$ and $g(y)$, $G(y)$ are the p.d.f. and c.d.f. of x and y respectively, then the c.d.f. of $u = x+y$ is given by

$$H(u) = \text{Prob}\,(x+y \leqslant u) = \int_0^\infty \int_0^{u-y} f(x)\,g(y)\,\mathrm{d}x\,\mathrm{d}y$$

$$= \int_0^\infty F(u-y)\,g(y)\,\mathrm{d}y.$$

Differentiation gives the p.d.f. as

$$h(u) = \int_0^u f(u-y)\,g(y)\,\mathrm{d}y$$

Similarly

$$h(u) = \int_0^u f(x)\,g(u-x)\,\mathrm{d}x.$$

We say that h is the convolution of f and g, and write

$$h = f * g.$$

It can be shown that

$$\mathscr{L}(h) = \mathscr{L}(f)\,.\,\mathscr{L}(g).$$

For example, if single gaps x are all distributed according to the negative exponential $\lambda e^{-\lambda x}$, the distribution of double gaps $u = 2x$ is given by the p.d.f.

$$(\lambda e^{-\lambda x})^{2*} = \int_0^u \lambda e^{-\lambda x}\,.\,\lambda e^{-\lambda(u-x)}\,\mathrm{d}x$$

$$= \lambda^2 u\, e^{-\lambda u}.$$

Now, with this notation, let the major road headways be negative exponential so that

$$f(t) = \mu e^{-\mu t}, \quad F(t) = 1 - e^{-\mu t}, \quad t \geqslant 0.$$

Then, by integration, immediately

$$\phi(s) = \mu/(\mu+s).$$

Let a pedestrian arrive at time $t = 0$, and require time T to cross so that the pedestrian delay has p.d.f. $w(t)$, c.d.f. $W(t)$ and L.T. $\psi(s)$.

If the first gap is $>T$ there is no delay and hence $w(t)$ has a discrete component at the origin of height

$$\text{Prob.}\,(t_1 > T) = 1-F(\mathrm{T}) = \mathrm{e}^{-\mu T}.$$

In addition there is a continuous component in the range $(0, \infty)$.

If the pedestrian must wait for gap t_{n+1}, say, where $n > 0$, then the delay is the sum of n random variables

$$\sum_{i=1}^{n} t_i.$$

Now the pedestrian has been unable to cross in any of the intervals $t_1, \ldots, t_n$, and all these are $<T$. Hence the density functions of all the t_i, $i \leqslant n$ are not simply $f(t)$ but $f(t)$ normalized over the range $(0, T)$, i.e.

$$\frac{f(t)}{F(T)}, \quad 0 < t < T$$

$$= \frac{\mu\,\mathrm{e}^{-\mu t}}{1-\mathrm{e}^{-\mu T}}.$$

Hence the continuous portion of $w(t)$ conditional on the crossing being in the $(n+1)$ *st* gap t_{n+1} is given by the convolution

$$\left[\frac{f(t)}{F(T)}\right]^{*}\left[\frac{f(t)}{F(T)}\right]^{\overline{n-1}*}.$$

Multiplying by the probability of crossing in the $(n+1)$ *st* gap, viz. $F^n(T)\ [1-F(T)]$, to obtain the unconditional component of $w(t)$, summing over n and adding the discrete component

$$w(t) = [1-F(t)]\,\delta(t) + [1-F(T)] \sum_{n=1}^{\infty} [f(t)]^{*}[f(t)]^{\overline{n-1}*}, \text{ where } f^{0*} = 1.$$

Hence, taking the Laplace Transform of both sides

$$\psi(s) = [1-F(T)] + \frac{[1-F(T)] \int\limits_0^T f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t}{1-\int\limits_0^T f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t}.$$

For the special case of the random distribution this turns out to be

$$\psi(s) = e^{-\mu T} + \frac{\mu e^{-\mu T}(1-e^{-(\mu+s)T})}{(\mu+s)\left[1-\dfrac{\mu}{\mu+s}(1-e^{-(\mu+s)T})\right]}$$

$$= \frac{(s+\mu)\,e^{-\mu T}}{s+\mu\,e^{-(\mu+s)T}}.$$

This function is somewhat difficult to invert but the solution can be checked to be

$$w(t) = e^{-\mu T}\,\delta(t) + \mu\,e^{-\mu T}\sum_{j=0}^{r-1}(-e^{-\mu T})^j\left\{\frac{[\mu(t-jT)]^{j-1}}{(j-1)!}+\frac{[\mu(t-jT)]^j}{(j)!}\right\},$$

$$(r-1)T \leqslant t \leqslant rT, \quad r = 1, 2, \ldots.$$

If $j=0$, the expression in { } is zero and then

$$w(t) = e^{-\mu T}.$$

In any case we need not actually carry out the inversion in order to obtain the mean and variance. We need only differentiate $\psi(s)$ to give the mean delay to all pedestrians thus

$$\text{mean delay} = -\left\{\frac{d}{ds}\psi(s)\right\}_{s=0} = \frac{1}{\mu}(e^{\mu T}-\mu T-1).$$

The variance similarly is given by

$$\left\{\frac{d^2}{ds^2}\psi(s)-\left[\frac{d}{ds}\psi(s)\right]^2\right\}_{s=0},$$

which after some reduction is found to be

$$\frac{1}{\mu^2}(e^{2\mu T}-2\mu T\,e^{\mu T}-1).$$

Omitting the discrete component the mean delay to all delayed pedestrians is found in a similar manner to be

$$\frac{1}{\mu}e^{\mu T}-\frac{T}{1-e^{-\mu T}}.$$

Raff (1951) has dealt with a similar problem involving traffic at an intersection controlled by a stop sign. He used a different approach which is outlined here.

In this model, the vehicles on the main road pass the intersection in a Poisson process given by

$$P(k \text{ vehicles in } t) = \frac{e^{-\mu t}(\mu t)^k}{k!}, \quad k = 0, 1, 2, \ldots$$

where μ is the mean number of arrivals in unit time. The arrival-time of each minor road vehicle is assumed arbitrary; but the time at which such a vehicle enters the intersection depends on the time-gap between its arrival and the passage of the next main-road vehicle. If this quantity is longer than some fixed quantity T, the minor road vehicle proceeds without delay. Otherwise it waits until a vehicle-free gap longer than T becomes available. From the point of view of minor-road vehicles the main road traffic is an alternate succession of blocks and 'anti-blocks'. All the time which precedes the passage of any vehicle by T or less is contained in blocks, whilst the rest of the time is contained in antiblocks, *vide* Fig. 8.2.

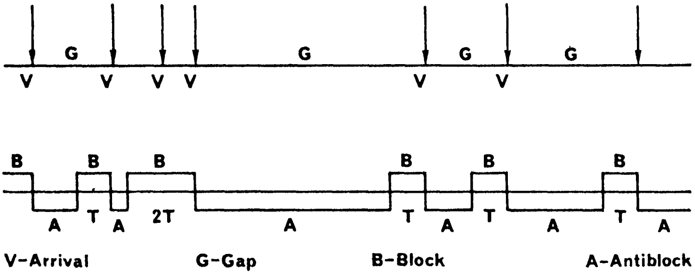

FIG. 8.2. Diagram showing position of blocks and anti-blocks in a stream of traffic.

We want to find (i) the probability that an instant selected at random is contained in a block, and (ii) the probability distribution of block sizes, i.e. the probability that a block, selected at random from the whole set of blocks, will have its size in a preassigned range.

The probability distribution of intervals between arrivals of vehicles, i.e. gaps, is given by the negative exponential, $f(t) = \mu e^{-\mu t}$, $t \geqslant 0$, where μ is the average number of vehicles per unit time. Then the probability that a gap selected at random is less than t is

$$F(t) = 1 - e^{-\mu t}. \tag{8.1}$$

Let the total number of vehicles passing in a long finite period of time be N. Then, the total time is N/μ. The first problem posed above in (i) can then be expressed otherwise thus: what fraction of the total time is contained in blocks? Since the total time is known, we then have the total time taken up by both blocks and antiblocks.

Now each antiblock is part of a gap $>T$. Conversely in every gap of length $T+t$, $t>0$, the first t is antiblock. This means that there is a (1, 1) correspondence between antiblocks and gaps $>T$. The number of antiblocks $>t$ is the same as the number of gaps $>(T+t)$ which, from equation (8.1) is

$$N\,\mathrm{e}^{-\mu(T+t)}.$$

The total number of antiblocks is $N\,\mathrm{e}^{-\mu T}$, hence

the number of antiblocks/unit time $=\mu\,\mathrm{e}^{-\mu T}$

and the number of antiblocks $\leqslant t = N\,\mathrm{e}^{-\mu T}(1-\mathrm{e}^{-\mu t})$. (8.2)

Now the total time in antiblocks is obtained by multiplying the derivative of equation (8.2) by t and integrating from 0 to ∞, i.e.

$$\begin{aligned}\mu N\mathrm{e}^{-\mu T}\int_0^\infty t\,\mathrm{e}^{-\mu t}\,\mathrm{d}t &= N\mu\,\mathrm{e}^{-\mu T}\left\{\left[\frac{t\,\mathrm{e}^{-\mu t}}{-\mu}\right]_0^\infty + \frac{1}{\mu}\int_0^\infty \mathrm{e}^{-\mu t}\,\mathrm{d}t\right\}\\ &= \frac{N\,\mathrm{e}^{-\mu T}}{\mu}\,[\mathrm{e}^{-\mu t}]_0^\infty\\ &= \frac{N}{\mu}\,\mathrm{e}^{-\mu T},\end{aligned}$$

so that the probability that a random instant is contained in an antiblock is $\mathrm{e}^{-\mu T}$, and the probability that a random instant is contained in a block is $(1-\mathrm{e}^{-\mu T})$.

The second problem posed above is more difficult because of the absence of a simple relationship between a particular block and any gap or group of gaps. However, it is possible to approach the problem indirectly through the distribution of waiting times for the minor-road vehicles; this latter distribution is closely related to the one we want.

Let $B(t)$ be the probability that a block $\leqslant t$. Let $W(t)$ be the probability that the waiting time of a minor road vehicle is $\leqslant t$. Then $W(0)=\mathrm{e}^{-\mu T}$ as before, since a vehicle arriving in an antiblock has zero waiting time. Now the probability that the waiting time t is in the range $u<t\leqslant u+\mathrm{d}u$ is the same as the probability that an instant is in

a block, and also precedes the end of the block by an amount of time which is between u and $u + \mathrm{d}u$. This is equal to $\mathrm{d}u$ multiplied by the average number of blocks/unit time which are longer than u, i.e.

$$\mu \, \mathrm{e}^{-\mu T}[1 - B(u)] \, \mathrm{d}u.$$

Thus

$$W(t) = W(0) + \mu \, \mathrm{e}^{-\mu T} \int_0^t [1 - B(u)] \, \mathrm{d}u$$

$$= \mathrm{e}^{-\mu T}(1 + \mu t) - \mu \, \mathrm{e}^{-\mu T} \int_0^t B(u) \, \mathrm{d}u. \qquad (8.3)$$

In order to express $B(t)$ in terms of $W(t)$ it is only necessary to differentiate both sides of equation (8.3) with respect to t and solve the resulting equation for $B(t)$, thus

$$W'(t) = \mu \, \mathrm{e}^{-\mu T}[1 - B(t)]$$

or

$$B(t) = 1 - \frac{1}{\mu \, \mathrm{e}^{-\mu T}} W'(t). \qquad (8.4)$$

The cumulative distribution of waiting times $W(t)$ follows directly from a formula developed by Garwood (1940)

$$W(t) = \mathrm{e}^{-\mu T} \sum_{j=0}^{r-1} (-\mathrm{e}^{-\mu T})^j \left[\frac{[\mu(t - jT)]^j}{j!} + \frac{[\mu(t - jT)]^{j+1}}{(j+1)!} \right],$$

$$(r-1)T \leqslant t \leqslant rT.$$

Garwood developed this formula from consideration of a problem connected with traffic lights which was mathematically almost the same as this, and later he applied the results to pedestrian delays for one or two traffic streams. His derivation proceeded from an expression for the probability distribution of the longest interval in a line of unit length divided by n random points.

Garwood's formula for $W(t)$ consists of a finite series in which the number of terms increases with increasing values of t, and it is a function of μT and t/T only, where μT is the traffic volume and t/T is the waiting time, the unit in both cases being the critical lag T. The quantity $W(t)$ has been computed for various values of μT and t/T but for large values of the parameters an approximation can be used:

$$W(t) \to 1 - (1 - \mathrm{e}^{-\mu T}) \exp\left[\frac{-\mu T \, \mathrm{e}^{-\mu T}(1 - \mathrm{e}^{-\mu T}) \, t}{(1 - \mathrm{e}^{-\mu T} - \mu T \, \mathrm{e}^{-\mu T}) \, T} \right]. \qquad (8.5)$$

The error involved in using this formula for most practical values of the parameters is of the order of 10^{-2}.

The following points should be noted. $W(t)$ increases linearly as t goes from 0 to T, and then there is a discontinuity in the slope. $W(t)$ decreases monotonically as μT increases, i.e. with increasing traffic volume, as one would expect. Differentiation of Garwood's formula for $W(t)$ gives the p.d.f. $w(t)$ directly thus

$$w(t) = W'(t) = \mu e^{-\mu T}[1-W(t-T)]. \tag{8.6}$$

This shows that $W'(t)$ is a linear function of $W(t-T)$. A comparison of equations (8.4) and (8.6) gives the c.d.f. of block sizes as

$$B(t) = W(t-T). \tag{8.7}$$

Thus the average block length is greater than the average waiting time by an amount T.

Some other properties of this distribution are worthy of note. There are no blocks $<T$, i.e. $B(t)=0$ for $t<T$. This is obvious from geometrical considerations. Also there are a certain number of blocks exactly equal to T, i.e. $B(t)$ jumps from 0 to $e^{-\mu T}$ at $t = T$. This is also apparent geometrically since a block of length T occurs whenever there are two consecutive gaps $>T$.

Since tht total time in blocks is $(N/\mu)(1-e^{-\mu T})$ and the total number of blocks is $Ne^{-\mu T}$, the average (expected) value of block length is

$$\frac{1-e^{-\mu T}}{\mu e^{-\mu T}}.$$

Suppose, as an example, we take the average rate of flow of vehicles per hour as 1,200 so that $\mu=1{,}200/3{,}600$, with a typical value for T of 6 sec. The proportion of blocks $\leqslant 15$ sec is given by

$$B(15) = W(15-6) = W(9).$$

Then substituting

$\mu T = 2$ and $t/T = 1{\cdot}5$ in equation (8.5), we get $W(9) = 0{\cdot}514$.

The above theory can be applied to the problem of pedestrian delays, with the following results:

(i) The proportion of pedestrians delayed $= (1-e^{-\mu T})$.

(ii) The average delay to delayed pedestrians $=$

$$1/(\mu e^{-\mu T})-T/(1-e^{-\mu T}).$$

(iii) The average delay to all pedestrians =

$$1/(\mu e^{-\mu T}) - 1/\mu - T.$$

and the variance =

$$(e^{2\mu T} - 2\mu T e^{\mu T} - 1)/\mu^2.$$

Table 8.1 gives some values of the delay to pedestrians for various values of the average flow μ and critical lag T.

TABLE 8.1

T	Flow v.p.h.	Proportion of pedestrians delayed.	Average delay to all pedestrians.	Average delay to delayed pedestrians.
4	100	0·1052	0·2300	2·2072
	200	0·1992	0·4784	2·3981
	300	0·2834	0·7472	2·6329
	400	0·3588	1·0364	2·8881
	500	0·4263	1·3496	3·1665
	600	0·4866	1·6868	3·4665
	700	0·5406	2·0522	3·7959
	800	0·5889	2·4463	4·1540
	900	0·6321	2·8732	4·5450
	1,000	0·6708	3·3357	4·9727
8	100	0·1992	0·9568	4·7962
	200	0·3588	2·0728	5·7763
	300	0·4866	3·3736	6·9330
	400	0·5889	4·8925	8·3079
	500	0·6708	6·6736	9·9475
	600	0·7364	8·7634	11·8997
	700	0·7889	11·2236	14·2258
	800	0·8310	14·1256	16·9986
	900	0·8647	17·5564	20·3046
	1,000	0·8916	21·6194	24·2468

In order to compare theoretical delays with observation we equate the following quantities:

p/P and $e^{-\mu T}$;

D/P and $1/(\mu e^{-\mu T}) - 1/\mu - T$;

$D/(P-p)$ and $1/(\mu e^{-\mu T}) - T/(1 - e^{-\mu T})$;

where μ = number of vehicles/sec,
P = total number of pedestrians,
p = number of pedestrians able to cross without delay,
D = total delay sustained by all pedestrians.

Analyses similar to the above have been carried out by a number of authors including Tanner (1951), who also derived distributions for the sizes of groups of pedestrians waiting to cross a road, assuming random arrivals for both vehicles and pedestrians.

Let α = number of pedestrians waiting to cross when a random vehicle passes;
β = number of pedestrians crossing immediately behind a random vehicle;
γ = size of a random group crossing;
δ = number of pedestrians waiting at a random time;
ϵ = size of group with which a random pedestrian crosses;
n = mean number of pedestrians arriving/unit time;
μ = mean number of vehicles arriving/unit time;
T = critical gap.

Writing E for the expected or mean value, Tanner's results are:

$$\text{(i)}\quad E(\alpha)=\frac{n(e^{\mu T}-1)}{\mu};$$

$$\text{(ii)}\quad E(\beta)=\frac{n(1-e^{-\mu T})}{\mu};$$

$$\text{(iii)}\quad E(\gamma)=\frac{\mu e^{-\mu T}+n e^{-nT}}{(\mu+n)e^{(n-\mu)T}};$$

$$\text{(iv)}\quad E(\delta)=\frac{n(e^{\mu T}-\mu T-1)}{\mu};$$

$$\text{(v)}\quad E(\epsilon)=1+\frac{2n}{\mu}(e^{\mu T}-\mu T-1);$$

We end this section by mentioning a matter of some direct practical importance. The point at which a traffic engineer decides to canalize pedestrians onto a crossing is usually taken as the point at which the traffic flows at such a rate as to cause 50% of all pedestrians to suffer a mean delay of 2 sec, and this corresponds to a mean delay to delayed pedestrians of about 4 sec, and to a traffic volume of about 130 passenger car units/h.

8.3 More Complicated Models

Tanner (1962) has considered a T-junction model which involves a number of assumptions but which is fairly realistic, and has given a formula for the average delay to vehicles on the minor road when the system is in statistical equilibrium, i.e. when vehicles could, in the long run, pass through the junction at a greater rate than they in fact arrive. The formula is derived for random flow on both major and minor roads.

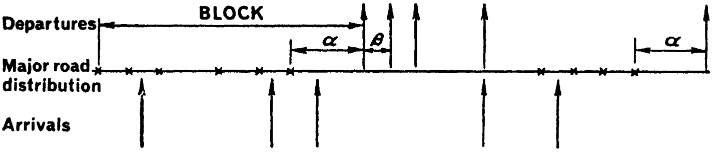

FIG. 8.3. Diagram showing arrivals and departures at a T-junction.

The mathematical model is as follows. The major road traffic consists of a single stream of vehicles which arrives at the junction at random at an average rate q_1/unit time, but which cannot pass through the junction at shorter intervals than a time β_1. Thus the vehicles on the major road form a queueing process and the average delay per vehicle is, according to queueing theory.

$$\bar{w}_1 = \frac{\frac{1}{2}\beta_1^2 q_1}{1-\beta_1 q_1}.$$

Vehicles on the minor road also arrive at random but at a rate q_2/unit time and β_2 is the shortest interval for passage through the intersection. Further, they cannot enter the junction within a time $\alpha > \beta_1$ after the previous road vehicle. There is no restriction on how closely a major road vehicle can follow a minor road vehicle.

On the major road, therefore, there is a series of blocks and gaps. Within the blocks the intervals of separation between the vehicles are less than α, and a gap is greater than α. During the gap, minor road vehicles pass onto the major road at intervals β_2 until the next block occurs, as shown in Fig. 8.3. Note that if $\beta_1 = 0$, traffic on the major road is random; the purpose of the bottleneck is to introduce an element of bunching into the major road traffic.

For this model Tanner derived a formula for the average delay $\overline{w}_2$ to vehicles on the minor road when the system is in statistical equilibrium.

Since arrivals on the major road are random, the distribution of the duration of a gap is negative exponential i.e.

$$f(t) = q_1 e^{-q_1 t}, \quad 0 < t < \infty. \tag{8.8}$$

The m.g.f. of the distribution of the duration of a block is given by Tanner (1953) as

$$E(e^{pt}) = M(t) = \frac{\nu(q_1 - p)}{\nu q_1 - (\nu q_1 - q_1 + p)\exp[(q_1 - p)(\alpha - \beta_1)]}, \tag{8.9}$$

where ν is the smaller real root of

$$\nu = e^{\beta_1 q_1(\nu - 1) + \beta_1 p}$$

and p is a parameter. This relation then gives

$$E(t) = \frac{e^{q_1(\alpha - \beta_1)}}{q_1(1 - \beta_1 q_1)} - \frac{1}{q_1}, \tag{8.10}$$

and

$$\tfrac{1}{2}E(t^2) = \frac{e^{q_1(\alpha-\beta_1)}}{q_1^2(1 - \beta_1 q_1)^2}\Bigg[e^{q_1(\alpha-\beta_1)} - \alpha q_1(1 - \beta_1 q_1) - 1 + \beta_1 q_1 - \beta_1^2 q_1^2 + \frac{\tfrac{1}{2}\beta_1^2 q_1^2}{1 - \beta_1 q_1}\Bigg]. \tag{8.11}$$

The method of regeneration points, as described by Kendall (1951) was used by Tanner in the determination of $\overline{w}_2$, which he gives as

$$\overline{w}_2 = \frac{E(t^2)/2y - q_2 y(\beta_2 q_1 e^{-\beta_2 q_1} + e^{-\beta_2 q_1} - 1)/q_1}{1 - q_2 y(1 - e^{-\beta_2 q_1}} \tag{8.12}$$

where $y = E(t) + 1/q_1$.

For the random model, $E(t)$ and $E(t^2)$ are given by (8.10) and (8.11) but (8.12) is valid for any distribution for which the method of regeneration points can be used.

We note, first of all, some special cases. If $q_1 = 0$ the minor road vehicles form a queue with uninterrupted service and (8.12) reduces to the usual formula

$$\overline{w}_2 = \frac{\tfrac{1}{2}\beta_2^2 q_2}{1 - \beta_2 q_2}.$$

If either q_2 or $\beta_2 = 0$ minor road vehicles do not obstruct each other and (8.12) reduces to

$$\overline{w}_2 = \frac{e^{q_1(\alpha-\beta_1)}}{q_1(1-\beta_1 q_1)} - \alpha - \frac{(1-\beta_1 q_1+\beta_1^2 q_1^2)}{q_1(1-\beta_1 q_1)} + \frac{\frac{1}{2}\beta_1^2 q_1}{(1-\beta_1 q_1)^2}.$$

Tanner also noted that $\overline{w}_2 \to \infty$ as

$$q_2 \to q_2(\max) = \frac{q_1(1-\beta_1 q_1)}{e^{q_1(\alpha-\beta_1)}(1-e^{-\beta_2 q_1})}.$$

This is the greatest flow on the minor road under the conditions assumed in the model. The sum of $q_2(\max)$ and q_1 gives the ultimate capacity of the intersection for a particular flow rate.

Tabulations of the delay $\overline{w}_2$ were carried out by Tanner for eighty different combinations of α, β_1, q_1, β_2 and q_2. If we take as an example

$$\alpha = 8 \text{ sec};$$

$$\beta_1 = 1 \text{ sec};$$

$$\beta_2 = 3 \text{ sec};$$

and if the flows on the major and minor roads are 900 v.p.h. and 72 v.p.h. respectively

then $$q_1 = 900/3{,}600 = 0{\cdot}3$$

and $$q_2 = 72/3{,}600 = 0{\cdot}02.$$

This means that 2 vehicles need $(8+3) = 11$ sec to cross, 3 vehicles need $(8+6) = 14$ sec and so on. Substituting in (8.12)

$$w_2 \approx 28 \text{ sec}.$$

A more realistic model would involve the minor-road vehicle crossing two major road streams, q_1 and q_2, one in each direction. This would require knowledge of the distributions of blocks and gaps on the major road when the two opposing streams are regarded as one. The gaps in the combined stream are negative exponential with mean $1/(q_1+q_2)$, but the distribution of block lengths does not appear to have been derived to date.

Miller (1963) has considered delays to pedestrians or minor-road vehicles for the case where the traffic on the major road occurs in bunches or queues, as described in §7.1. In this model he postulated queues of vehicles randomly distributed on the road, with the queues arriving at a rate λ and the flow of queues being μ. The mean waiting time for a vehicle on the minor road is given as

$$E(t) = \frac{\lambda(\bar{t}+\tau)^2}{2}$$

where $\bar{t}$ = average time for a queue to pass a point,

τ = minimum acceptable gap.

The probability that a minor road vehicle will be able to cross immediately is

$$P(0) = (1-\mu\bar{t})\,e^{-\lambda\tau},$$

where $1/\mu = \bar{t}+1/\lambda$.

To take an example if $\tau = 5$, $\bar{t} = 10$, and the flow rate of queues is 120/h, i.e. $\lambda = 120/3{,}600 = 1/30$, then

$$E(t) = \frac{(10+5)^2}{60}$$

$$= 3{\cdot}75 \text{ sec.}$$

Also $\mu = 1/40$ so that

$$P(0) = \left[1-\frac{10}{40}\right]e^{-0{\cdot}33}$$

$$= 0{\cdot}54.$$

Miller compared his model with that of the random vehicles model. He found little difference in the mean waiting time for vehicles wishing to cross the main stream. He concluded that the random bunches model, however, predicted opportunities for crossing without delay better than the random vehicles model.

None of the models considered in this section fully accounts for what is often the most significant cause of delay for vehicles crossing the main stream, i.e. the additional delay due to waiting in a queue behind other vehicles. Weiss and Maradudin (1962) have considered the case where two vehicles arrive simultaneously, but theoretical treatment of a more realistic model is quite difficult.

The validity of some of the models is open to question. T-junctions corresponding to the conditions of the models are not always easy to find. Major road flow must be uninterrupted in most cases; this eliminates all intersections where right-turning vehicles from the minor road may hold up the flow. Since minor road flow is also taken to be uninterrupted, the models do not allow pedestrians to cross the intersection nor vehicles to turn into the minor road from the major road.

It is also true that drivers do not behave altogether systematically in the T-junction situation. For instance, the major road flows may be interrupted by a courteous driver slowing down to allow a minor road vehicle to emerge. Another version of this is where minor road vehicles accept too small a gap, causing major road vehicles to slow down.

More realistic models are unlikely to be susceptible to analytical methods, and recourse will probably have to be had to simulation, an approach to the solution of traffic problems which is described in Chapter 9. As has already been mentioned the criterion for performance at an intersection has usually been taken to be the average delay to a minor road vehicle. This quantity is of mathematical interest, but under present-day conditions is of secondary importance compared with that of capacity. This quantity is, in fact, easier to study, both empirically and theoretically, and there seems to be better prospects of relating the two approaches.

8.4 Signal-Controlled Intersections

A large number of authors have dealt with delays at fixed-cycle traffic lights, but most of the models are simple and idealized because of the mathematical difficulties involved. We shall describe first some of the results due to Little (1961) for certain idealized situations, and then pass on to some approximate methods due to Newell (1965) for dealing with more complex and realistic situations.

Little has formulated a number of models, and from them has predicted the delay to vehicles in a number of situations. In order to make the models sufficiently practical only the most important variables were included, and as a consequence the results are applicable only for low to medium flows.

The models considered include that of queueing at a traffic signal. Arrivals are taken as random and in a single lane. At the signals the traffic is stopped for a time T and then released, the vehicles being allowed to proceed only straight ahead. It is assumed that stopping and starting are instantaneous, and that starting vehicles have a constant headway h between them. This assumption about instantaneous accelerations is not serious in practice as the extra delay involved can be partly eliminated by using an 'effective' red equal to the actual red plus the average delay due to acceleration.

For low to medium traffic flows, i.e. for the case where there is no carry-over from one red to the next, Little developed an equation for the average queue-length at a traffic signal. This is

$$E(N) = \frac{qR}{1-qh}$$

where q = flow in vehicles/sec/lane,

R = effective red time,

h = time-headway at the signal.

If h is assumed constant, the average time for the queue to pass through the intersection is

$$E(t) = E(N)h.$$

This model can be extended to the case of multiple traffic lanes for two particular cases. In the first model arriving vehicles join the shortest queue at the signal. If n is the number of approach lanes the average time for the queue to be dissipated is given approximately by

$$E(t) = E(N)\frac{h}{n}.$$

The second model seems to give a better approximation. In this, arrivals form n separate and independent streams of traffic. If the maximum queue-length in any stream is Q, the average time for a queue to pass is Qh. If each lane is considered as a separate stream the average length of queue in each lane is

$$E(N) = \frac{(q/n)\,R}{1-(q/n)\,h} = \frac{qR}{n-qh}$$

where q is still the total flow.

To take an example, if the total flow is 600 v.p.h. over two lanes, the effective red time is 60 sec and h is 3 sec then

$$q = \frac{600}{2.3600} = \frac{1}{12} = 0{\cdot}083 \text{ veh/sec},$$

and
$$E(N) = \frac{0{\cdot}083(60)}{1-0{\cdot}083(3)} \approx 7.$$

Other authors who have studied the signalized intersection include Hawkes (1963), Newell (1960) and Darroch (1964). Newell has studied the discrete time queue at a traffic light for the special case where at most one vehicle arrives per unit time. Darroch has extended this work to include a general arrival process.

In most papers output flow has been taken as uniform during the green phase and input as simply stochastic. However, when signals are in close proximity to one another, input has a well-defined cyclic character. This is particularly true of linked systems. Blunden and Pretty (1965) used cyclic output and input flows, and represented them as continuous variables by means of Fourier series.

Real signal devices and real traffic are so complicated that an exact analysis of a realistic model is impossible. For this reason some authors have used fluid analogies instead. The simplest of these is due to Clayton (1941). In a more recent paper Newell (1965) has dealt with the problem by using approximations based upon a representation of the queue as a continuous fluid with either deterministic or stochastic properties. These approximations are based on the use of a law of large numbers (or a central limit theorem) and are not very sensitive to the form taken for the arrival or departure process. The results for the delays in the applications considered by Newell are correct to within a few per cent. The approximations employed are, in fact, asymptotic and apply strictly only in the case where the average queue-length is infinite. In practice they could be used for an average queue-length of, say, 10.

In Clayton's model vehicles arrive at regular intervals (headway $1/q$), form a queue during the red time R of the traffic signal and leave during the succeeding green, again at regular intervals (headway $1/s$). This process continues until either the end of the green or the end of the queue, whichever is earlier. It is easy to evaluate the queue-lengths at any time for this model but there are some algebraic complications due to the fact that the quantities $1/q$, $1/s$, R, G may not be rational multiples of one another. These complications disappear if the flow of vehicles is thought of as a continuous fluid arriving at rate q per unit time, and released after a time R at rate s per unit time in the first instance then at rate q.

Let $A(t)$, $D(t)$ be the cumulative number of arrivals and departures respectively in time t of the green phase when the queue is non-empty, so that

$$A(t) = qt, \quad D(t) = st.$$

Let $Q(t)$ be the queue-length at time t and take the start of a red phase as $t=0$. Then $Q(t)$ is given by

$$Q(t) = \begin{cases} Q(0)+A(t), & 0 < t \leqslant R \\ Q(0)+A(t)-D(t-R), & R \leqslant t \leqslant t_0 \\ 0, & t_0 \leqslant t \leqslant R+G \end{cases} \tag{8.13}$$

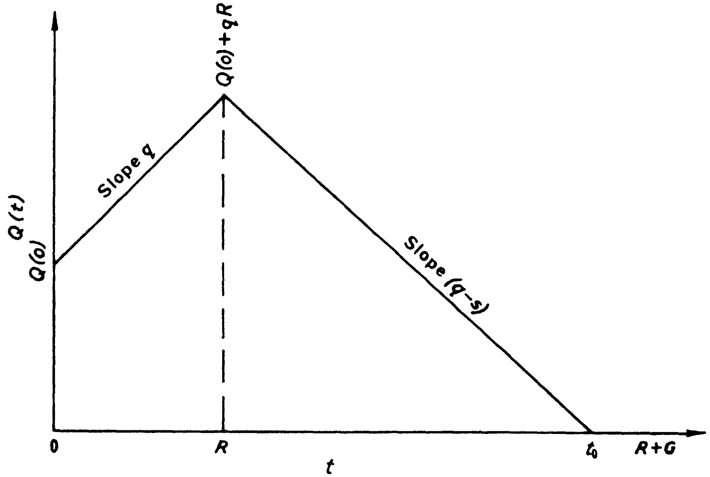

FIG. 8.4. Diagram showing queue-length at a signalized intersection.

where t_0 is the time when the queue first vanishes. $Q(t)$ takes the form shown in Fig. 8.4. Since the curve is piecewise continuous, t_0 is given by

$$Q(0)+A(t_0)+D(t_0-R) = 0$$

i.e. $$t_0 = \frac{Q(0)+qR}{s-q}+R \quad \text{for} \quad t_0 < R+G.$$

If $t_0 > R+G$ the equation for $Q(t)$ still applies so long as $t \leqslant R+G$.

In order to find an expression for the total delay note that the total delay to all vehicles in $(t, t+\delta t)$ is $Q(t)\,\delta t$, and hence in time $(R+G)$ it is the area under the curve given by equation (8.13), i.e.

$$W = \int_0^{R+G} Q(t)\,\mathrm{d}t.$$

For $t_0 \leqslant R+G$

$$W = [RQ(0)+\tfrac{1}{2}qR^2) + \frac{[Q(0)+qR]^2}{2(s-q)}.$$

For $t_0 \geqslant R+G$

$$W = GQ(R+G)+RQ(0)+\tfrac{1}{2}qR^2+\tfrac{1}{2}(s-q)\,G^2.$$

Since the queue at $(R+G)$ is given by

$$Q(R+G) = Q(0)+qR-(s-q)\,G,$$

the latter expression reduces to

$$W = (R+G)\,Q(0)+qRG-\tfrac{1}{2}(s-q)\,G^2+\tfrac{1}{2}qR^2.$$

Hence

$$W = \begin{array}{ll} [RQ(0)+\tfrac{1}{2}qR^2]+\dfrac{[Q(0)+qR]^2}{2(s-q)}, & t_0 \leqslant R+G \\ (R+G)\,Q(0)+qRG-\tfrac{1}{2}(s-q)\,G^2+\tfrac{1}{2}qR^2, & t_0 \geqslant R+G. \end{array} \tag{8.14}$$

These crude formulae can be applied to a fixed-cycle traffic light with alternating red and green periods. If $q(R+G) < sG$ and $Q(0)$ is finite, an equilibrium state will eventually be reached in which $Q(t)$ will become zero at the start of some red and $Q(t)$ will be periodic. In this case the total delay is given by (8.14) with $Q(0)=0$,

i.e.
$$W = \frac{sqR^2}{2(s-q)}. \tag{8.15}$$

Then the average delay per vehicle is

$$w = \frac{W}{q(R+G)} = \frac{R^2}{2(R+G)(1-q/s)}.$$

For $q(R+G) > sG$ the queue grows and the delay per cycle becomes infinite. For two intersecting streams the red time for one stream is equal to the green plus amber of the other, and the delay to both streams per cycle is obtained as the sum of two expressions like (8.15) with appropriate values for R.

Newell has generalized Clayton's approach and used asymptotic approximations to simplify the results in a way similar to their use in the theory of Brownian motion. In practice, R, G are of comparable size and the saturation flow per cycle sG is of the order of 10. However, let us assume first that $sG \to \infty$. This means that either $R \to \infty$ and $G \to \infty$ with R/G fixed or $q \to \infty$ and $s \to \infty$ with times measured in units of $(R+G)$ i.e. R and G of $0(1)$. If arrivals or departures form a stationary stochastic process the numbers of arrivals or departures should satisfy a law of large numbers. This means that for any limiting process in which the average number of arrivals $\to \infty$ then if N is the random number of arrivals, $N/E(N) \to 1$ and we have a continuous fluid. A sufficient (but not necessary) condition for this is that the intervals between arrivals are independent, in which case a renewal process is defined.

Newell suggested that $A(t)$, $D(t)$ should be regarded as random variables such that

$$E[A(t)] = qt, \quad E[D(t)] = st. \tag{8.16}$$

$$\text{Then as } sG \to \infty, \quad \frac{A(t)}{sG}, \to \frac{qt}{sG}, \ \frac{D(t)}{sG} \to \frac{t}{G}$$

with probability unity. The other formulae hold in the same sense if

$Q(t)$ is measured in units of sG,
t_0 is measured in units of G,
W is measured in units of sG^2,
w is measured in units of G.

Practical situations do not usually conform to these conditions. However, if we are primarily interested in total delays, and if $Q(t)$ is some random function of the time, it is not important for equations (8.16) to hold for every T. It is only necessary that the area under the curve defined by $Q(t)$ has small fluctuations. If the queue-length fluctuations cannot be neglected we need the expectations of queue-lengths, i.e.

$$E(W) = \int_0^{R+G} E[Q(t)]\,\mathrm{d}t.$$

Over a long period, n cycles say, the number of arrivals is $nq(R+G)$ approximately, and the total waiting time is $nq(R+G).E(W)$. This is equal to $nE(W)$, hence, for almost any reasonable stationary arrival pattern

$$E(w) = \frac{E(W)}{q(R+G)},$$

Now the total waiting time W is made up of waiting during the red, W_R, and waiting during the green, W_G. But

$$\begin{aligned} E(W_R) &= \int_0^R E[Q(0)+A(t)]\,\mathrm{d}t \\ &= RE[Q(0)]+\int_0^R qt\,\mathrm{d}t \end{aligned}$$

hence
$$E(W) = RE[Q(0)]+\tfrac{1}{2}qR^2+E(W_G). \tag{8.17}$$

If q/s, R/G are small compared with 1, the average time taken for the queue to clear will be of the order qR/s, which is small compared with G. The probability that in equilibrium the queue will not be discharged during G should be extremely small, and also $E(W_G)$ should be small compared with $E(W_R)$. There are cases where $E(W_G)$ cannot be neglected, but since there are complications in the evaluation of this quantity, we shall consider only the case where it can be ignored. Hence, except for unusual arrival patterns, for example batches of order sG,

$$E(W) \sim \tfrac{1}{2}qR^2 \quad \text{for} \quad q/s \ll 1,$$

and this can be compared with the expression (8.15). Note that this formula is now true for any arrival pattern and does not require long queues or cycle-times. In fact it is exact as $q \to 0$ and $E[Q(t)] \to 0$.

If $sG \gg 1$ there is a range of values of q for which $E(W_G)$ is not negligible but where q is still sufficiently small for there to be a negligible probability of the queue failing to clear during G, i.e. $RE[Q(0)]$ can still be neglected in equation (8.17). The additional term in $E(W)$ can be shown to be of the order of $1/sG$, except for nearly saturated flow in which there is a carry-over from one red to the next, hence

$$E(W) = \frac{qR^2}{2(1-q/s)} \text{ to } 0(1/sG)$$

and
$$E\,(w) = \frac{R^2}{2(R+G)(1-q/s)}.$$

For flows close to saturation a better approximation to $E(W)$ is given by

$$E(W) = \frac{qR^2}{2(1-q/s)} + (R+G)E[Q(0)].$$

Using methods similar to those used in the theory of Brownian motion, Newell obtained an approximation to $E[Q(0)]$, thus

$$E[Q(0)] = \frac{Iq}{2s}\left[\frac{G}{R+G} - \frac{q}{s}\right]^{-1},$$

where I is the sum of the coefficients of variance (ratio of variance to mean) of the cumulative distributions of arrivals and departures. This is of $0(1/sG)$ except for saturation flow. Using this expression for $E[Q(0)]$ we get the expected value of delay per vehicle as

$$E(w) = \frac{R^2}{2(R+G)(1-q/s)} + \frac{I}{2s}\left[\frac{G}{R+G} - \frac{q}{s}\right]^{-1}. \tag{8.18}$$

For values of $sG \sim 10$, which are those of interest in practice, $1/sG$ is unfortunately not really small. Errors are therefore of the order of 10%, but these are just about tolerable in the context of traffic measurements.

It is interesting to compare the formula for $E(w)$ given by (8.18) with Webster's formula (6.1). In the notation of this section this is

$$E(w) = \frac{R^2}{2(R+G)(1-q/s)} + \frac{x}{2s}\left[\frac{G}{R+G} - \frac{q}{s}\right]^{-1} - 0{\cdot}65\frac{(R+G)}{(sG)^{2/3}}x^{\alpha}, \tag{8.19}$$

where $x = q(R+G)/sG =$ degree of saturation,

$$\alpha = \frac{4}{3} + \frac{5G}{R+G}.$$

The first terms of (8.18) and (8.19) are identical. The second terms differ only in Webster's formula having x where (8.18) has I. In the computer simulation Webster used Poisson arrivals so that here $I_A = 1$. Also I_D is small, hence we can take $I \approx 1$ and then the two terms agree at least for $x = 1$ or saturation flow. Webster's third term was an empirical correction. The two formulae agree numerically to within the expected 10–15% accuracy.

8.5 Delays on an Open Road

The problem of delays on a two-lane road has been discussed by a number of authors including Tanner (1961) and Yeo (1964). In this problem the quantities of interest are mean journey-speed and the distributions of intervals and queue-lengths. Road transport undertakings, for instance, are interested in the intervals between consecutive buses on a given route. For a simple model Tanner derived a theoretical formula for the mean speed maintained by a vehicle on a long journey. He also suggested extensions to the model which, at the present time, are probably beyond the scope of mathematical analysis. Yeo has also generalized Tanner's simple model.

This model postulates a long straight road having two lanes, one in each direction. Except for the vehicle studied, all vehicles in one direction travel at the same speed and are spaced at random with a certain minimum distance between the vehicles. Similarly for the opposing traffic but with different speeds and spacing. The vehicle studied catches up and overtakes a bunch of vehicles at minimum spacing in a single manoeuvre. It overtakes without delay if there is a large enough gap in the opposing traffic, and the minimum acceptable gap increases linearly with the size of the bunch being overtaken. If a delay occurs, then a correspondingly greater gap is required in order to allow for the time lost in accelerating.

Numerical values of the mean speed were tabulated by Tanner, taking various values for the constants of the model. Two general conclusions were indicated by the numerical results. Firstly, more acceleration, obtained at the expense of reduced maximum speed would probably provide increased mean journey speeds under typical traffic conditions. Secondly, the cutting down of clearances when overtaking would not usually provide any wothwhile increase in journey speed. It was noted also that when the traffic flows increased beyond a certain level, which was appreciably below the theoretical capacity of the road, the fast vehicle could not maintain a higher mean speed than that of the other vehicles: nearly the whole time was spent waiting for opportunities to overtake. The mean waiting time in fact becomes infinite.

Tanner has suggested a number of extensions to this model. The most obvious one would be to allow for a stream of faster vehicles, instead of only one. These faster vehicles could be assumed to have identical properties.

Another line of development would be to study non-stationary behaviour. If the fast vehicle were injected at a random point into the traffic stream, in what way would its expected speed decrease to the mean speed already given by the analysis of the simple model? Tanner suggested that the initial expected speed would decrease only very slowly to this mean speed, and in this case most journeys would not be long enough for the latter to be a useful approximation.

Problems relating to interaction between lanes of traffic are virtually untouched as yet. Not even the grossest qualitative features have been discussed, nor does there seem to be any empirical work in this field.

8.6 Bottlenecks

Let us consider again the fluid analogy of Lighthill and Whitham. There is no indication in this theory that there are any circumstances which could lead to the flow at a bottleneck falling below its normal capacity. This phenomenon does exist in road traffic, and it cannot be accounted for by a simple kinematic wave theory. It is known that the introduction of short gaps in a traffic stream helps to increase the flow at a bottleneck. One explanation of this is that shock-waves are thereby eliminated. Or it may be connected with the fact that accelerations and decelerations of vehicles have finite limits.

This latter fact is one reason why the fluid flow model is imperfect. Other reasons have been given in §4.2. A more fundamental and detailed investigation of traffic behaviour at a bottleneck, therefore, is necessary for a complete understanding of all the phenomena.

Most investigators have tried to avoid this approach, because of the statistical difficulties involved in it. It has been suggested by Welding (1963) that these difficulties might be overcome by using time series analysis. This is a quite different approach to the problem, involving as it does a technique for the analysis of empirical data.

Consider a single line of traffic passing through a bottleneck. If the rate of input of vehicles is gradually increased, then as the flow approaches the capacity of the bottleneck, the speed of the vehicles in the bottleneck will decrease, and if the flow exceeds the capacity a queue will start to form. A vehicle passing through this system experiences four definite states viz. free flow before the bottleneck, queueing, flow in the bottleneck and free flow after the bottleneck. Thus it appears that the best characteristic to use for detecting the position of the bottleneck is vehicle speed; the bottleneck occurs at the point where traffic begins to accelerate.

In order to make use of time series analysis the first step is to take three points on the road, above, in and below the bottleneck and, for these points, record the passage-times of the vehicles and their instantaneous speeds. This could be done by means of some sort of automatic device.

Then this data could be processed to form a number of indices in an attempt to deduce how the design of the bottleneck was affecting its capacity, and also to indicate whether wave motion was affecting maximum flow.

The mean speed could be used, as indicated above, to pinpoint the exact location of the bottleneck. The capacity of the bottleneck could be indicated by the mean interval at a series of points, and its variation as the input flow is increased and a queue begins to form.

If $i(n)$ is the time-interval between the nth and $(n+1)$th vehicles and $i(m)$ is the mean interval, the quantity

$$c_r = E[i(n)-i(m)][i(n+r)-i(m)]$$

for various r could be used to indicate whether wave phenomena in the bottleneck were reducing its capacity. Similarly the quantities

$$c_{\sigma\tau} = E[i_s(t)-i_s(m)][i_{s+\sigma}(t+\tau)-i_{s+\sigma}(m)],$$

and

$$c'_{\sigma\tau} = E[v_s(t)-v_s(m)][v_{s+\sigma}(t+\tau)-v_{s+\sigma}(m)],$$

for various σ τ where $i_s(t)$ and $v_s(t)$ are the time-interval in the traffic stream and the speed of the vehicle respectively at point s and time t, could be used to investigate the wave motion on the approach to or exit from the bottleneck. Peaks in the curve for $c_{\sigma\tau}$ or $c'_{\sigma\tau}$ at (σ', τ') say, would indicate a corresponding wave velocity of σ'/τ'.

En passant one might mention that the above quantities, although they look somewhat formidable, can readily be calculated on a computer. Ideally the computer could read the data from a tape produced directly by the road detectors. This approach might give us valuable information which would throw light on such questions as whether the geometrical layout of a bottleneck affects its behaviour, or whether, and in what circumstances, the formation of wave motion aggravates the situation.

CHAPTER 9

Simulation of Traffic Problems

9.1 Introduction

The first step towards the solution of a scientific problem is always its precise formulation. After that there are a number of possibilities. The methods of solution can be divided into three main types: analytical methods, numerical methods and simulation. Analytical methods of different kinds using stochastic processes, differential equations and so on have already been described in the preceding chapters. Numerical methods for the solution of such things as differential or integral equations may be necessary at some stage, and very often analytical and numerical procedures are used together. Simulation offers a quite different method of solution and may be carried out on either an analogue or on a digital machine. There are advantages and disadvantages in both cases.

Usually an analogue or continuous-variable machine is set up to solve some specific problem, and within the set program may be very efficient. The computation, in this case, is carried out by varying the state of certain physical elements. Its accuracy is limited by that of the elements themselves, the relative error being usually at least of the order of 10^{-5}.

A digital or discrete-variable computer is, in general, a more flexible instrument which can be switched to a wide variety of problems, Results obtained by using a general-purpose computer—the 'software' approach – can be valuable, but the job can be speeded up by using a machine built specially to study and simulate problems of traffic flow. This is the hardware method. Such a machine 'The Road Traffic Simulator' has been built in the Digital Techniques Laboratory of the Electrical Engineering Department of the University of Manchester. It has been used to study the flow at intersections controlled both by traffic lights and by roundabouts where a right-hand priority rule operates.

All these types have their importance, but we shall be concerned here only with the simulation of traffic problems using software (programs) and a general-purpose digital computer. Some knowledge about the working of these computers will be assumed.

Digital simulation has been defined as 'the technique of setting up a stochastic model of a real system which neither over-simplifies the system to the point where the model becomes trivial, nor incorporates so many features of the real system that the model becomes intractable or phohibitively clumsy'. It is about 10 years since the first use was made of these techniques on a computer, although simulation of industrial processes, engineering structures and so on by means of Gantt charts and scale models, has been in use for many years. Since then the flexibility of the digital computer in this field has been put beyond question.

A simulation study using a computer usually involves the following steps. The scientific problem is first defined, and then a precise and detailed model formulated. Block diagrams may be of some use here. This model must include a clear statement of the assumptions, and the selection of a criterion for measuring the effectiveness of the system. Then follows the preparation of the computer program, and possibly some experimental runs to determine the values of the parameters. Finally the results of the computer output must be interpreted and evaluated.

The simulation approach is particularly useful when the situation is so complex that a mathematical solution is difficult or impossible to obtain, at least within a reasonable time. The special value of simulation lies in the ease with which parameters can be varied. The validity of the models must always be checked against data collected in real situations. Then statistics, experimentation and simulation may produce valid results.

It has been found that comparatively rough models will give results that agree reasonably well with both observation and theory. However, usable results have not always come out of these studies, for various reasons. One of the main handicaps is the lack of proper mathematical models, and the fact that over-simplification of the model may limit its validity, or the study may be too specific to be of general value. Lack of computer capacity may also be a cause. There are also problems of a different kind: problems of education and communication, which may cause the highway engineer to be sceptical about the results and make him unwilling to use them for design purposes, for example.

Highway traffic simulation was first investigated in 1954 by Mathewson, Trautman and Gerlough. They considered a 4-arm signal-controlled intersection with one lane in each direction. Pedestrians were present and both right and left turns were permitted, but right turns were only permitted on the green signal. They showed that such an intersection could be simulated by a general-purpose discrete-variable model, but it was not until the following year that such a model was constructed and yielded results.

This model, due to Goode, Pollmar and Wright (1956) was intended as a first step in the creation of a model to simulate a series of intersections, but in fact it got no further than a model for a single intersection of a simple type. The idea was to allow for the representation of individual vehicles in the model, the movement of the vehicles to be determined according to the conditions surrounding the vehicle at any particular moment. It was based on streets 22 ft wide with input lanes of approximately 400 ft in length, and the vehicles were taken as having an average length of 18 ft. They travelled in lanes at 30 m.p.h. when unobstructed, passed through an intersection with three-phase traffic lights and they turned right, left or proceeded straight ahead according to what were called 'driver desires'. Various assumptions were made, such as no overtaking and no interference from parked cars or pedestrians. Entry into each of the four lanes was controlled by a pseudo-random number routine. The traffic lights were simulated by cycle instructions, and the frequencies of the various turning movements were governed by a stochastic mechanism which compared a generated random number with the probability of each turning movement.

There are two possible ways of representing this problem. In the first model a number recorded in the computer represents the vehicle's present position, its direction of motion, its speed, its desired destination and so on. Scans are carried out at short intervals and vehicles are 'moved' by operations on the model. The second model – easier to program but less flexible in operation – was the one chosen. In this case the geographical area is broken up into blocks of approximately one vehicle-length. The model is scanned at short intervals in order to determine whether or not each position is occupied by a vehicle. If the position is occupied then the situation is handled according to certain rules.

The model, when constructed, was used to simulate an intersection under various conditions – different flows, different percentages of turning vehicles, different phase times for the traffic lights etc. – and the average delays at the traffic signals calculated. The model could obviously be extended by making the output from one set of signals the input for another. The simulation was very slow, in fact three times as slow as real time !

D. L. Gerlough (1957) also examined two methods of approaching the simulation model: he considered both a physical and a mathematical representation. In the former the roadway is considered to be made up of several storage cells laid end to end, and the presence or absence of vehicles are represented by 1's and 0's respectively. Forward movement of the vehicles is accomplished by any mathematical operation which causes 'bits' to move in a forward direction within any given cell or across the boundary into the next cell. In the latter a number or 'word' is assigned to represent each vehicle. Various groups of digits within the word are assigned to represent various items such as location, speed, length and so on. In this representation a vehicle is moved forward by simply adding the quantity (speed, time, etc.) to the appropriate 'word'. Gerlough concluded that the method of mathematical representation offered the most promise.

A more recently reported simulation model was that constructed by the Road Research Laboratory and reported by J. C. Benson (1962). This was a model of a 4-arm, 2-phase intersection having vehicle-actuated traffic signals. The variation of the average delay to vehicles with different settings for the traffic signals and under various conditions of flow were thoroughly investigated. The model was general and could be adapted easily to fit different types of signals and any intersection of this type.

Let us now consider a typical simulation in more detail. We must first consider the factors which must be included in the simulation, remembering that it is unnecessary and even unwise to include factors which contribute little to the experiment just for the sake of realism. We wish, then, to construct a mathematical model which contains the essential features but excludes irrelevances. We should remember, too, that the computer model can sometimes be considerably reduced in size by the careful introduction of analytical techniques.

The units of traffic involved are vehicles (with drivers) and pedestrians, and the movement of these units involves considerations of speed, volume, road capacity and the concentration and distribution of vehicles. Out of these we expect to get results about delays, journey-times and so on.

The characteristics of traffic units to be included in the model may include some or all of the following:

(i) Drivers make decisions at variable time intervals, dictated by changes in the traffic situation.
(ii) Drivers have a finite, non-zero reaction time.
(iii) Vehicles have finite acceleration and deceleration limits, and maximum speeds.
(iv) Vehicle speed is governed by speed regulations, driver temperament etc. as well as by factors mentioned under (iii).
(v) Vehicles are characterized by their size, power and various mechanical properties.
(vi) Drivers (or pedestrians) are affected by psychological traits, fatigue, the weather, and so on.
(vii) Drivers (or pedestrians) have limited ability to judge headway distances and vehicle speeds.
(viii) In overtaking or junction situations (drivers), or pedestrian crossing situations (pedestrians) there are definite probabilities that a driver (pedestrian) will or will not accept a gap in the oncoming traffic.

The characteristics of traffic movement include:

(i) Traffic speed, and hence the distances covered.
(ii) Traffic concentration or density – this determines the longitudinal distribution of vehicles along the road.
(iii) Traffic volume.
(iv) Road capacity.
(v) Lateral distribution of vehicles, i.e. number of lanes in each direction – this governs the possibility of merging, overtaking and so on.
(vi) Parked vehicles or obstructions, which, in particular, may affect flow through an intersection.

We also make various assumptions about the 'normality' of the driver, the 'averageness' of the vehicle and the 'neutrality' of the weather.

Suppose that a simulation is being carried out on a network of roads. The procedure is usually something as follows. The simulation is carried out in discrete time-intervals, called scan intervals, of t seconds. Vehicles move along the links between the signals at free flow speed unless they are impeded, in which case they stop. During each scanning cycle four major operations are performed, as follows:

(i) Vehicles are generated on the input links.
(ii) Vehicles are moved from links to connecting links.
(iii) Vehicles are advanced from zone to zone within the links.
(iv) Traffic-signal states are updated.

Each vehicle in the network gives rise to a movement or a delay, and these are accumulated to give the various traffic statistics.

The results obtained from a simulation problem on an electronic computer can be obtained in a variety of forms. For example, on the one hand one can obtain the printed output of all the detailed occurrences such as arrival times and departure times, or, on the other hand, one can obtain distribution functions in the form of histograms of queue-length, for example. These would be obtained by sampling the length of the queue at specified time intervals and accumulating the results.

9.2 A Queueing Model for a T-junction

In two papers Evans, Herman & Weiss (1963, 1964) have described a simulation investigation involving a T-junction. The headways on the main road are assumed to be distributed exponentially with arbitrary mean headway and the minor road arrivals are Poisson with arbitrary mean arrival rate. The gap acceptance functions employed are either a step function, a trapezoidal function or a delayed exponential with arbitrary parameters. The action of a waiting vehicle to merge or cross the highway is made dependent on a logical decision. Thus it is arranged that the leading vehicle in the queue should accept the gap γ with probability $p(\gamma)$, determined from the gap acceptance function being used and the selection of a number from a random number routine, programmed in the electronic computer. If the leading vehicle fails to accept a gap, the outcome is noted and the gap acceptance process is repated in full at the beginning of the next gap. On the other hand, if the leading vehicles does accept the gap, the second vehicle in the queue is first shifted to the leading position with a move-up time γ', and then the above procedure is repeated with the diminished gap $(\gamma-\gamma')$.

All the vehicles in the queue are assumed to shift up to the next position in the queue simultaneously. In the paper of Evans *et al.* the move-up time γ' is treated as a constant but in more general cases it would be taken as a random variable. It should be noted that because of the difficulty in handling all the factors explicitly, the move-up time associated with each driver-vehicle system includes not only the time required by the movement, but also decision times and all other relevant factors taken together.

We have already shown in §7.3 that when the traffic intensity is small – in other words when the average rate of arrival on the minor road is small compared with the average rate of arrival on the main road – the queue at the stop sign will go to zero infinitely often with probability unity. On the other hand, if the traffic intensity approaches unity – in other words if the average arrival rate on the minor road is large compared with that on the main road – the queue on the minor road will, on the average, grow steadily with probability unity. There is a critical point, which can be calculated theoretically, for which all queue-lengths are equally probable, and for this reason the queue-length observations fluctuate wildly. It can be shown that the critical value for the arrival rate on the minor road is given by λ_c, where

$$\lambda_c = \frac{\mu\, e^{-\mu T}}{1-e^{-\mu\gamma'}} \tag{9.1}$$

and μ and λ are the average rates of arrival on the major road and minor road respectively, γ' is the constant move-up time for the vehicles in the queue and T is the minimum gap acceptance time for a step acceptance function.

If $\lambda > \lambda_c$ the queue is transient, if $\lambda < \lambda_c$ the queue is in equilibrium. In the latter case it is possible to find P_0, the probability that there are no vehicles in the queue. When t is large the expected number of vehicles on the minor-road is λt exactly, and the expected number discharged from the junction is

$$\lambda_c(1-P_0)t+0(1).$$

If an equilibrium state exists as $t \to \infty$

$$\lambda t = \lambda_c(1-P_0)t,$$

i.e.

$$P_0 = 1-\frac{\lambda}{\lambda_c}.$$

This gives a method for evaluating λ_c experimentally, using Monte-Carlo techniques.

Note that in the limit as $\mu \to 0$ in equation (9.1), $\lambda_c \to 1/\gamma'$, and the move-up time becomes important. The actual gap-acceptance function used, however, can be shown to be not critical.

Evans *et al.* considered the following specific numerical problem for which queue-length histograms can be obtained from theory. Let the arrival rates on both major and minor roads be random with $\mu = 0{\cdot}2$ vehicles/sec and λ variable. Let the minimum gap acceptance for the step function and the move-up time be equal so the $T = \gamma' = 7$sec.

In this case

$$\begin{aligned}\lambda_c &= \frac{\mu}{e^{\mu T}-1} \\ &= \frac{0{\cdot}2}{e^{1{\cdot}4}-1} \\ &= 0{\cdot}06546.\end{aligned}$$

If λ/λ_c is taken equal to 75%, 90% and 95% of λ_c respectively, i.e. if $\lambda = 0{\cdot}050, 0{\cdot}060$ and $0{\cdot}063$, the average length of queue for each case can be obtained from the formula

$$E(n) = \frac{\lambda}{\mu}\frac{[(\lambda+\mu)(e^{\mu T}-1)-\lambda T e^{\mu T}]}{[(1-e^{\mu T})\lambda+\mu]}. \qquad (9.2)$$

Note that if λ is set equal to λ_c the denominator of equation (9.2) vanishes so that the average queue-length is infinite. The values of $E(n)$ as calculated from equation (9.2) for the three values of λ given above are found to be 2·5, 8·16 and 18·7. Evans *et al.* compared these with the results of four simulation runs of 4,000 vehicles in each case. Agreement with the theoretical average queue-length is quite good, but the scatter of the values for the individual runs varies considerably. The scatter is large for $\lambda/\lambda_c = 95\%$, somewhat less for $\lambda/\lambda_c = 90\%$ and smaller still for $\lambda/\lambda_c = 75\%$.

We note again that for values of λ close to λ_c the average length of the queue is large and the length of the queue as a function of the time varies very markedly. In addition we know from queueing theory that in the limit, $\lambda = \lambda_c$, every queue-length occurs (in the long run) with equal probability. Obviously for a finite simulation this cannot be the case; what must happen is that for long, but finite simulations the histograms of queue-length will vary considerably from run to run. The occurrence of two long bunches of vehicles on the major-road can produce appreciable queueing times, and yet long runs with no queueing can occur. The estimated mean queueing time may, therefore, have a serious random error, and also the behaviour of the real system taken over moderate time periods may depart appreciably from that predicted by the equilibrium distribution.

For observations made in the field, in the neighbourhood of λ_c, we can obtain sensible information only in respect of certain specific questions. We cannot predict anything about queue-length or waiting-time, but we can discuss the gap-acceptance behaviour of vehicles. Simulation techniques like the one described can be of great value where observations in the field require very long times, since these are clearly not only difficult, but may even be essentially impossible, because the required conditions do not remain constant for a sufficient length of time. Care must however, be taken in posing the questions and in interpreting the results.

9.3 Sampling from a Distribution

In order to construct a simulation model with sufficient realism it is necessary to fill it with 'real' traffic. One method would be to collect a lot of data from traffic situations similar to those experienced in the model and then use this data. 'Real' traffic, however, conforms, at least approximately, to statistical distributions – the random distribution for instance. So an actual population can be replaced by its theoretical counterpart and a sample taken from this, by means of a random number table, for example.

A distribution of speeds which is a reasonable approximation to that actually observed in urban streets can be obtained as follows. As each vehicle is assigned, three random numbers from a uniform distribution are produced, all in the range $(+1, -1)$. If these are denoted by r_1, r_2, r_3 then a vehicle-speed is formed by taking

$$(\text{desired average speed}) + \tfrac{1}{3}(r_1 + r_2 + r_3) \times (\text{desired speed range}).$$

In general vehicle characteristics such as speed, minimum acceptable gap and so on can be generated by using a random number generator of some kind and probability tables. In order to sample from a distribution uniform in (a,b), a table of random digits is used to give a random decimal x in $(0,1)$ and then the sample value is $x(b-a)+a$. This method can also be used with a table of random Normal deviates, for example, to generate a Normal distribution of speeds. The standard Normal distribution gives a value of x and then the sample value is $\mu+\sigma x$.

A better procedure when using a computer is to use the inverse cumulative distribution function, described below, since rapid means are available in this case for generating (pseudo-)random numbers uniformly distributed over the range $(0,1)$.

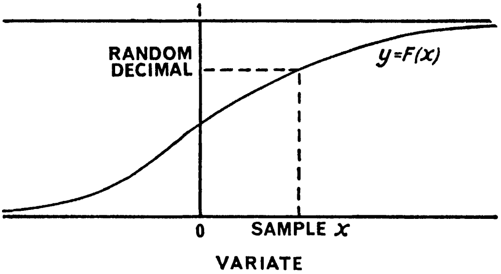

FIG. 9.1. Cumulative probability $F(x)$ plotted against the variate x.

If the cumulative probability $F(x)$, or $1-F(x)$, is plotted against the values of the variate x we get a graph which looks like the one in Fig. 9.1. The cumulative probability gives the probability of obtaining the given value of the variate or less from the sampling distribution. In order to select an item the procedure is to choose a random decimal between 0 and 1, i.e. a point on the vertical axis, and then the corresponding value on the horizontal axis is the sample value. This is equivalent analytically to finding the inverse cumulative probability.

In order to justify this procedure it is necessary to show that any value in the population has the same chance of being chosen as any other or, in other words, that the probability of getting a sample value between x_1 and $x_1+\mathrm{d}x$ is proportional to $f(x_1)$. This follows immediately since

$$\text{Prob}\,[x_1 < \text{sample}\, x < x_1+\mathrm{d}x) = F'(x_1)\,\mathrm{d}x = f(x_1)$$

as required.

A number of ways of obtaining the inverse cumulative distribution function are possible, viz. tabular inversion; numerical approximation; Monte Carlo or trial and error procedures; analytic inversion of the c.d.f. and the calculation of this function for a selected uniform random variate.

A number of authors have discussed these procedures, among them Tocher (1963) in his book *The Art of Simulation.* He discusses also a number of other topics, such as random and pseudo-random numbers, which are used in simulation studies. We shall give here only three methods for sampling from three of the most important distributions, the Poisson, the negative exponential and the Normal.

For the Poisson distribution we proceed as follows:

(i) Generate a random number R, say.

(ii) Form the cumulative Poisson distribution

$$P_n = \sum_{r=0}^{n} \frac{\mu^r e^{-\mu}}{r!}, \quad \text{for } n = 0, 1, 2, \ldots.$$

(iii) Compare P_n with R at each step and take the first value of n such that $P(n) \geqslant R$ as a random variate.

The random variate thus generated gives the number of arrivals. This process can be carried out as and when required, or a table can be computed initially and table look-up used when required.

For the negative exponential the prodcedure is equally simple.

Here $$p(x) = \lambda e^{-\lambda x}, \quad P(x) = 1 - e^{-\lambda x}$$

and the inverse is

$$x = \frac{-1}{\lambda} \ln(1-P).$$

If P, $(1-P)$ are uniform in $(0,1)$ we have the rule, to find an exponential variate take

$$\frac{-1}{\lambda} \ln(\text{u.r.v.}),$$

where λ is a parameter, and u.r.v. is a uniform random variate.

A general analytic method has been suggested by a number of authors including Butcher (1961). This is substantially as follows. Suppose that $f_i(x)$, $i = 1 \ldots n$ are frequency functions for those distributions for which a sampling method is known, and that $f(x)$ is a frequency function for the distribution to be sampled. Suppose also that we can write

$$f(x) = \sum_{i=1}^{n} \alpha_i f_i(x) g_i(x), \quad \alpha_i > 0, i = 1 \ldots n;$$
$$0 < g_i(x) < 1, i = 1 \ldots n.$$

The following procedure gives variates from the required distribution:

(i) Choose an integer, ν, in $(0, n)$ such that

$$\Pr(\nu = i) = \alpha_i/(\alpha_i + \ldots + \alpha_n).$$

(ii) Choose a value, x, from the distribution with frequency function $f_\nu(x)$.

(iii) Compute $g_\nu(x)$ and compare it with a random number u. Then if $u \leqslant g(x)$ the value of x chosen in (ii) is accepted as the required sample value, otherwise it is rejected and all the steps are repeated. Note that the average number of values of ν and x that must be chosen before a value x is accepted is $\sum_{i=1}^{n} \alpha_i$.

Consider, as an example, a standardized Normal distribution. Since it is a simple matter on most computers to allot a sign randomly, it is sufficient to consider the modulus of the variate only. The frequency function of the distribution to be sampled is then

$$f(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2}, \quad x > 0.$$

Applying the general procedure outlined above with $n = 1$ and $f_1(x) = \lambda e^{-\lambda x}$, for some λ, we get

$$\alpha_1 g_1(x) = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2 + \lambda x}$$
$$= \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}(x-\lambda)^2 + \frac{1}{2}\lambda^2}.$$

We can, therefore, choose

$$g_1(x) = e^{-\frac{1}{2}(x-\lambda)^2},$$

since this is always less than or equal to 1. In this case

$$\alpha_1 = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{\frac{1}{2}\lambda^2}.$$

A value of $\lambda = 1$ minimizes this quantity (and hence the time taken to apply the method, since the two are proportional), and hence

$$\alpha_1 = \sqrt{\frac{2\mathrm{e}}{\pi}} = 1{\cdot}315.$$

The condition for acceptance then becomes

$$u \leqslant \mathrm{e}^{-\frac{1}{2}(x-1)^2}$$

or

$$(x-1)^2 \leqslant -2\ln u$$
$$= 2y,$$

where y, like x, is sampled from the exponential distribution with parameter unity.

CHAPTER 10

Accident Statistics

10.1 Accident Proneness and other Problems

Problems of a quite different kind from those discussed in previous chapters also exist. Since the beginning of the century, road accidents in Britain have caused ten million casualties, of which 285,000 were deaths. In 1964 the figures were the worst since records began in 1909: there were 6,922 deaths out of a total of 356,179 casualties. Predictions for 1965 are about 8,000 deaths. If this figure is realized the year will rank with the years 1939, 1940 and 1941 when the figures were of comparable magnitude.

The subject of road accidents is obviously of great social importance, and it is increasingly a topic for scientific investigation by workers in a variety of fields. Since Greenwood and Yule published their paper in 1920 there have been literally hundreds of papers on various aspects of accident causation. However, few results of real scientific value have yet emerged. The reasons for this lie in the difficulties involved in accident investigations. These are twofold: in the first place there is the difficulty of getting non-corrupt data, and in the second place there are statistical difficulties.

In most work in the traffic field it is the liability of policies to claims rather than of drivers to accidents which has been studied. There are many complications here. Drivers are anxious not to lose their no-claim bonus and tend not to claim for accidents which result in minor damage only. The types of accidents involved are relevant, but the seriousness or otherwise of the results of an accident may have little connection with the causation. To categorize accidents by outcome could be valueless if the hypothesis is that the result is due to chance and nothing else. It might be argued with some justification that it is better to pool all kinds of accident, since a driver likely to incur one kind may also be likely to incur any other kind.

There are further snags in the investigation of accident statistics. The total mileage at risk is a highly variable quantity and it is, in addition, obviously correlated with other factors such as age and experience. Multiple correlation analysis is therefore indicated, though in most investigations known to the author it does not seem to have been used. Usually simple comparisons have been made of the effect of experience on accident rates for the different age groups separately. Where exposure to risk is not included as a variate it must be equalized over the population of drivers considered. This is difficult as it includes not only total mileage driven but also such factors as time of day, hours already spent driving, weather conditions, routes taken and so on.

In their book *The Causation of Bus Driver Accidents: an Epidemiological Study* Cresswell and Froggatt (1963) have shown that increasing experience acted as a significant factor in reducing the accident rate among drivers in the early years of driving. They found a high accident rate in the first year of driving, and the 18–23 age group had a particularly bad record. After the preliminary effects of experience had worn off, age appeared to have no effect up to about 55 years, then from 55 to 65 the accident rate increased. This effect must have been correctly attributed to age since the effect of increasing experience was shown to be in the reverse direction. Häkkinen (1958) obtained similar results for a population of bus and tram drivers in Helsinki. If these results are found to hold for the general population of drivers it could constitute a serious problem in accident prevention.

The most vexed question of all relates to the concept of 'accident proneness'. Is there such a thing? How do we define it? If it exists at all, is it a stable attribute? And is it general or quite specific, i.e. is an individual who is likely to incur an accident on the road also likely to incur one in the home or at work.

There is a certain amount of circumstantial evidence that there is such a thing as an innate tendency to accident which varies between individuals in the same way as, for example, characteristics such as physique or intelligence. It seems a reasonable hypothesis that individuals should vary in this way as they do in others, but the hypothesis has not been substantiated to date by any valid statistical evidence or argument. This does not mean that accident proneness does not exist: it merely means that is is not proven. A substantial literature exists on the psychological aspects of accident causation as a result of subconscious motivation. The interested reader is referred to it: further discussion of this aspect would be out of place here.

The term 'accident proneness' is widely used, but has never been defined in an entirely satisfactory manner. The original definition of Farmer and Chambers (1926) includes a distinction between the concepts of accident proneness and accident liability. They write, '"Accident proneness" is a narrower term than "accident liability" and means a personal iodiosyncrasy predisposing the individual who possesses it to a marked degree to a relatively high accident rate. "Accident liability" includes all the factors determining accident rate, "accident proneness" refers only to those that are personal.' This definition is to some extent ambiguous, but at least it is an improvement on the sort of circular definition which makes accident incidence *per se* a measure of accident proneness, for this is to adopt the position of those who say that accidents are due to carelessness, and when asked to define carelessness do so in such a way as to indicate that they mean having an undue number of accidents!

In most of these attempts at definition some distinction is made between factors personal to the individual and factors relating to his environment. Although it is obviously necessary to do this, there are some dangers in pursuing this distinction too far: it is the interaction between the individual and his environment which produces the end result (an accident) and it would be profitless to start in this context another individual-versus-environment controversy. Individual susceptibility is itself, of course, a term which covers a whole host of factors which are difficult to separate, let alone to measure.

Even if accident proneness exists there seems to be considerable doubt as to whether such an attribute could be stable. It seems more likely that it is a characteristic which varies throughout an individual's life and depends, at least to some extent, on such factors as health, home circumstances and so on. It is fairly certain, however, that the factor is specific to a certain situation (work, home, etc.) and is not a general attribute – in other words there is no such thing as a real 'calamity Joe'.

In §10.2 some of the theoretical distributions which have been used in the study of accident data are derived and their underlying hypotheses and their relative merits discussed. The reader is warned that a good many wild and misleading statements have been made in the literature, notwithstanding the fact that the earliest authors were perfectly aware of the pitfalls and pointed them out. Probably the most balanced recent account is that of Arbous and Kerrich (1951) which contains both the mathematical background of the subject and a critical evaluation of what it does and does not prove.

10.2 Theoretical Distributions

Historically the Poisson distribution was the first model to be used to formulate and express in terms of theoretical probabilities a specific hypothesis to explain the distribution of accidents amongst individuals exposed to an equal risk of incurring an accident in an unchanging environment. The Poisson distribution was discovered in 1837 but it was not until 1898 that it was used to describe the number of men killed by horse-kicks in a Prussian Army Corps over a period of 20 years. It came into further prominence when 'Student' (1907) showed that this distribution afforded an approximate model for a class of haemocytometer counts. The two classical papers in the accident field are those of Greenwood and Woods (1919) and of Greenwood and Yule (1920). Greenwood and Woods investigated the frequency of accidents among munition workers, and compared the observed frequency distribution with three theoretical distributions which they deduced from three different hypotheses. Greenwood and Yule, starting from the Poisson distribution, deduced a number of more sophisticated models to describe accident phenomena. Newbold (1926, 1927) critically examined and extended the work of these authors. More recently Cresswell and Froggatt have suggested two new hypotheses, and have derived the corresponding distributions. It should be noted that some of their statements appear to be suspect: for a critical review see Irwin (1964).

We shall now derive these distributions and discuss their application to accident statistics.

I *Poisson*

The first hypothesis to be considered was one of pure chance. This gave rise to the well-known Poisson distribution, already derived in §7.1, in which the probability of an individual having k accidents in a given time is

$$P(k) = \frac{e^{-\lambda}.\lambda^k}{k!}, \quad k = 0, 1, 2. \dots$$

where λ is the mean number of accidents/unit time.

In Britain the mean number of accidents per annum is in the range 0·2 to 2·0 for minor damage only, about 0·05 for accidents involving minor personal injury and 0·01 for serious accidents. These are the values of λ in the Poisson distribution for accidents.

We should like to emphasize a point which, although elementary, does not seem to be properly appreciated, viz. that the Poisson distribution does not give an odd result as some people seem to think. On the random hypothesis 12% of the drivers (say) have 45% of the accidents. In fact, the smaller the mean number of accidents (λ) – and it is often very small in accident statistics – the greater the apparent distortion. This is illustrated in Table 10.1 where λ is taken as 0·6.

TABLE 10.1

Number of accidents k	Proportion of individuals having k accidents		Proportion of accidents	
0	54·8		0·0	
1	33·0		33·0	
2	9·8	12%	19·6	45%
3	2·0		6·0	
4	0·4		1·6	
Total	100·0		60·2	

Notice that for $\lambda = 0{\cdot}6$ almost 50% of the population have at least one accident and more than 10% two or more. Even if an accident prone group exists, therefore, it could not be identified by considering only the number of accidents sustained by the individuals. It should also be emphasized that although one can argue that if the population at risk is homogeneous (i.e. if λ is constant throughout the population) then accident data relating to that population can be fitted by a Poisson distribution, the inverse deduction is *not* true.

II *Negative binomial*

Greenwood and Woods, analysing the incidence of accidents among munition workers during the first World War, observed that the Poisson distribution did not generally provide a suitable model, since the variance of the observed distribution was greater than the mean by an appreciable amount in most instances.

They retained the Poisson distribution as the distribution of accidents for a given λ but arranged for an *ab initio* differentiation among the population members, i.e. for λ to vary for different individuals. The conditional distribution is thus

$$P(k|\lambda) = \frac{e^{-\lambda}.\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The distribution of λ is one which should be skew to the right, and the one arbitrarily chosen was the gamma distribution because in practice it is easy to fit. This distribution is mathematically reasonable as a description of liability but is not of necessity a good one to choose. It is possible also to take here another Poisson distribution, in which case a Compound Poisson distribution results, as we shall see later.

The gamma distribution (*vide* §7.2) is given by

$$dF = \frac{c^p}{\Gamma(p)}.e^{-c\lambda}\lambda^{p-1}\,d\lambda, \quad 0 \leqslant \lambda < \infty, c > 0, p > 0.$$

The distribution which emerges from this hypothesis is then given by

$$P(k) = \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!}.\frac{c^p}{\Gamma(p)}e^{-c\lambda}\lambda^{p-1}\,d\lambda, \quad k = 0, 1, 2, \ldots$$

$$= \frac{c^p}{k!\Gamma(p)}\int_0^\infty \lambda^{k+p-1}e^{-\lambda(c+1)}\,d\lambda, \quad k = 0, 1, 2, \ldots$$

Putting $\mu = -(c+1)\lambda$, i.e. $d\mu = -(c+1)\,d\lambda$ we get

$$P(k) = \frac{c^p}{(c+1)^{k+p}k!\Gamma(p)}\int_0^\infty \mu^{k+p-1}e^{-\mu}\,d\mu$$

$$= \frac{\Gamma(k+p)}{k!\Gamma(p)}.\frac{c^p}{(c+1)^{k+p}}, \quad k = 0, 1, 2, \ldots$$

It can also be written as

$$P(k) = \left(\frac{c}{c+1}\right)^p {}^{k+p-1}C_{p-1}\frac{1}{(c+1)^k}, \quad k = 0, 1, 2, \ldots$$

$$= \left(\frac{c}{c+1}\right)^p {}^{-p}C_k\frac{(-1)^k}{(c+1)^k}, \quad k = 0, 1, 2, \ldots$$

This distribution is now known as the negative binomial: at that time it was known as the distribution of unequal liabilities. The hypothesis from which it has been derived here is the one corresponding to the accident proneness concept.

The mean and variance can be found from first principles or by differentiating the moment generating function w.r.t. t and putting $t=0$, thus

$$M_0(t) = \left(\frac{c}{c+1}\right)^p \left(1-\frac{\mathrm{e}^t}{c+1}\right)^{-p}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} M_0(t) = \frac{pc^p\,\mathrm{e}^t}{(c+1)^{p+1}}\left(1-\frac{\mathrm{e}^t}{c+1}\right)^{-p-1}$$

$$\left.\frac{\mathrm{d}}{\mathrm{d}t} M_0(t)\right|_{t=0} = \frac{p}{c}.$$

Hence the mean is p/c. Similarly

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_0(t) = \frac{pc^p\,\mathrm{e}^t}{(c+1)^{p+1}}\left(1-\frac{\mathrm{e}^t}{c+1}\right)^{-p-1} + \frac{p(p+1)\,c^p\,\mathrm{e}^{2t}}{(c+1)^{p+2}}\left(1-\frac{\mathrm{e}^t}{c+1}\right)^{-p-2}$$

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_0(t)\right|_{t=0} = \frac{p}{c}+\frac{p(p+1)}{c^2}$$

and hence the variance is

$$\left\{\frac{\mathrm{d}^2}{\mathrm{d}t^2} M_0(t) - \left[\frac{\mathrm{d}}{\mathrm{d}t} M_0(t)\right]^2\right\}_{t=0} = \frac{p}{c}+\frac{p(p+1}{c^2}-\frac{p^2}{c^2}$$

$$= \frac{p(c+1)}{c^2}.$$

Since p, c are both positive, the variance is always greater than the mean. In order to fit this distribution to observed data, the parameters p and c are found by equating the observed mean and variance to p/c and $p(c+1)/c^2$.

Greenwood and Yule found that this distribution fitted their accident data better than the Poisson, and this is generally true of accident data. But, as these authors pointed out, the closeness of fit is *not* proof that the underlying hypothesis is correct. The negative binomial can arise in a great many ways, and at least one other hypothesis is plausible in the accident context.

Suppose that the probability of sustaining an accident, originally the same for all individuals, is altered in some way after the first accident. One might argue that the occurrence of an accident might make an individual more conscious of the dangers of the situation and through the exercise of more care, less liable to accident in the future. In other words there would be a kind of learning process. One might conceivably argue the other way, and say that through nervousness or reduced physical skill, for example, an individual might be more predisposed to an accident than previously.

As early as 1914 McKendrick obtained the general solution for the case where the subjects exposed to risk start alike but after k accidents they have probability $f(k)$ of sustaining another. He pointed out that the negative binomial comes as a particular case of this, in which $f(k)$ is a linear increasing function of k. He derived the same distribution again in 1926.

Let $p(k,t)$ be the probability that an individual has had k accidents in time t, then assume that in time $(t, t+\delta t)$ he can have

$$1 \text{ accident with probability } f(k,t)\,\delta t,$$

$$0 \text{ accident with probability } [1-f(k,t)]\,\delta t,$$

and that the probability of two or more accidents in δt is a second order small quantity. Then by a process similar to that by which the Poisson distribution was derived in §7.1 we get

$$\left.\begin{aligned} p(k,t+\delta t) &= p(k,t)[1-f(k,t)]\,\delta t + p(k-1,t)f(k-1,t)\,\delta t + o(\delta t),\\ &\qquad\qquad\qquad\qquad\qquad\qquad k = 1,2,\ldots\\ p(0,t+\delta t) &= p(0,t)[1-f(0,t)]\,\delta t + o(\delta t).\end{aligned}\right\} \quad (10.1)$$

Subtracting $p(k,t)$ from both sides of equations (10.1), dividing by δt and taking the limit as $\delta t \to 0$ we have

$$\left.\begin{aligned} \frac{\partial}{\partial t}p(k,t) &= -p(k,t)f(k,t)+p(k-1,t)f(k-1,t), \quad k = 1,2,\ldots\\ \frac{\partial}{\partial t}p(0,t) &= -p(0,t)f(0,t).\end{aligned}\right\}$$

In the particular case where $f(k,t) = \beta + \gamma k$ for $\beta, \gamma > 0$ these equations become

$$\left.\begin{aligned} \frac{\partial}{\partial t}p(k,t) &= -p(k,t)[\beta+\gamma k]+p(k-1,t)[\beta+\gamma(k-1)], \quad k = 1,2,\ldots\\ \frac{\partial}{\partial t}p(0,t) &= -\beta p(0,t).\end{aligned}\right\}$$

i.e.

$$\left.\begin{aligned}\frac{\partial}{\partial t}p(k,t)+(\beta+\gamma k)p(k,t) &= [\beta+\gamma(k-1)]p(k-1,t),\\ k &= 1,2,\ldots\\ \frac{\partial}{\partial t}p(0,t)+\beta p(0,t) &= 0.\end{aligned}\right\}\quad(10.2)$$

The solution of these differential-difference equations can be obtained by generating function methods. If $G(k,s)$ is the generating function corresponding to $p(k,t)$ we have

$$G(k,s) = \sum_{0}^{\infty} s^k(\beta+\gamma k)p(k,t).$$

Multiplying the kth equation of the set (10.2) by s^k and summing over k we get

$$\frac{\partial}{\partial t}G(k,s) = (s-1)\left[\beta G(k,s)+\gamma s\frac{\partial}{\partial s}G(k,s)\right],$$

and the solution of this is

$$G(k,s) = [e^{\gamma t}-s(e^{\gamma t}-1)]^{-\beta/\gamma}.$$

At time $t=0$, $s=1$ and $G(k,s)=1$ hence

$$p(k,t) = \frac{\Gamma[\beta/\gamma+k]}{k!\Gamma(\beta/\gamma)}[1-e^{-\gamma t}]^k e^{-\beta t}.$$

This is a negative binomial distribution with

$$p = \beta/\gamma \quad\text{and}\quad c = (e^{\gamma t}-1)^{-1}.$$

The case $f(k,t)=\beta+\gamma k$ corresponds to an original constant probability β of an accident, followed by a linearly increasing liability. A decreasing liability would correspond to a function.

$$f(k,t) = \beta-\gamma k.$$

The resultant distribution in this case is the binomial. At first sight this looks reasonable, but note that for large k, $f(k,t)$ is negative and thus cannot represent a probability. The analysis would be invalid in this case.

To sum up, the negative binomial can arise on a number of hypotheses, at least two of which are plausible in the accident context. Quite apart from this, even where heterogeneity of the population is the cause, it need not be due to accident proneness: it could be due to differences of any sort in exposure to risk. In spite of this, the fact that a negative binomial often gives a closer fit than the Poisson to accident data has been used as a proof that accident proneness exists. In fact the most we are justified in saying is that if the negative binomial is a bad fit then *probably* this contradicts the accident proneness concept.

In practice it is necessary to determine the cause of accidents, as this necessarily determines the remedies to be applied. We must, therefore, be able to differentiate between the two hypotheses and for this purpose it is useless to consider only the univariate distribution, i.e. it is insufficient to consider only one period of exposure. On the other hand, if we consider a bivariate distribution, i.e. two periods of exposure, discrimination is possible since the two hypotheses lead to different bivariate distributions. These are discussed next.

III *Bivariate distributions*

If we consider two time periods and calculate the sample correlation coefficient (r_{01}) for the numbers of accidents in these two periods we may be able to make inferences about the population correlation coefficient (ρ_{01}). A zero value for ρ_{01} would indicate no correlation between the accident rates in the two time periods, and hence a random (Poisson) distribution for the accidents. If $\rho_{01} > 0$ this would indicate heterogeneity in the population, and if r_{01} is large enough it would be reasonable to assume that $\rho_{01} > 0$. This would lead to the tentative conclusion that over the whole period of time in question some drivers tend to have more accidents than others, and a bivariate analysis might help in isolating these individuals.

Let the total time interval $(0,T)$ be divided into two parts, $(0,t)$ and (t,T) and let these be called δ, δ_1, δ_2 respectively. Let the corresponding numbers of accidents in these intervals be k, k_1 and k_2 so that $k = (k_1 + k_2)$.

For the case of the bivariate negative binomial we start from the distributions

$$p(k_i|\lambda) = \frac{e^{-\lambda\delta_i}(\lambda\delta_i)^{k_i}}{k_i!}, \quad i = 1, 2$$

and

$$p(\lambda) = \frac{c^p}{\Gamma(p)} e^{-c\lambda}\lambda^{p-1}\,d\lambda.$$

It can be deduced, by moment generating function methods for example (see Arbous and Kerrich), that the 'marginal' distributions for k_1, k_2 and k are all negative binomial, thus

$$p(k) = \frac{c}{(c+\delta)} \frac{\Gamma(p+k)}{\Gamma(p)\,k!} \left(\frac{\delta}{c+\delta}\right)^k,$$

and similarly for $p(k_1)$ and $p(k_2)$.

The bivariate distribution for k_1 and k_2 is the bivariate negative binomial

$$p(k_1, k_2) = \frac{c^p}{(c+\delta)^{p+k}} \frac{\Gamma(p+k)\,\delta_1^{k_1}\,\delta_2^{k_2}}{\Gamma(p)\,k_1!\,k_2!}.$$

From these distributions we can find the means μ_1' and variances μ_2 etc., for instance

$$\mu_1'(k) = \frac{p}{c}\delta;$$

$$\mu_2(k) = \frac{p}{c}\delta\,(1+\delta/c);$$

$$\rho(k_1, k_2) = \frac{1}{\sqrt{(1+c/\delta_1)(1+c/\delta_2)}};$$

For the case of the bivariate contagious distribution the corresponding distributions and formulae are as follows:

$$p(k) = \frac{\mathrm{e}^{-\beta\delta}\,\Gamma(\beta/\gamma+k)}{k!\,\Gamma(\beta/\gamma)}(1-\mathrm{e}^{-\gamma\delta})^k,$$

and similarly for $p(k_1)$ and $p(k_2)$;

$$p(k_1, k_2) = \mathrm{e}^{-\beta\delta} \frac{\Gamma(\beta/\gamma+k)}{\Gamma(\beta/\gamma)\,k_1!\,k_2!}(\mathrm{e}^{-\gamma\delta_2}-\mathrm{e}^{-\gamma\delta})^{k_1}(1-\mathrm{e}^{-\gamma\delta_2})^{k_2};$$

$$\mu_1'(k) = \frac{\beta}{\gamma}(\mathrm{e}^{\gamma\delta}-1);$$

$$\mu_2(k) = \frac{\beta}{\gamma}(\mathrm{e}^{\gamma\delta}-1)\,\mathrm{e}^{\gamma\delta};$$

$$\rho(k_1, k_2) = \frac{1}{\sqrt{\left[1+\frac{1}{(\mathrm{e}^{\gamma\delta_1}-1)}\right]\left[1+\frac{1}{\mathrm{e}^{\gamma\delta_1}(\mathrm{e}^{\gamma\delta_2}-1)}\right]}}.$$

The sample means and variances given here can be used to obtain estimates of the parameters p and c or β and γ. Then the theoretical probabilities (or frequencies) corresponding to the marginal and bivariate distributions for both hypotheses can be calculated. As we have already seen, the values of $p(k)$ are the same on both hypotheses. The values in the two-way tables are different however. A χ^2 test can be used to compare the two theoretical distributions with the observed values and thus a decision made as to which is the better fit, and hence which is the more likely hypothesis.

IV *Long and Short distributions*

Two other distributions have been suggested by Cresswell and Froggatt which successfully reproduce observed accident data. They are based on the concept of an individual as a creature of altering efficiency, and indeed subject to periods of substandard efficiency, called, for short, 'spells'. These spells might be associated with temporary factors such as ill-health. Unlike the theory of accident proneness the model allows for the occurrence of accidents which are primarily due to other road users.

Suppose that an individual is subject to spells, and no accident can occur outside a spell, which is taken to be a rare event and one which occurs by chance, i.e. the number of spells in any one period is independent of the number in any other period. Further let us suppose that all drivers are equally likely to have a spell and that the chance of an accident in a spell is constant and independent of the chance of a spell.

Let λ be the mean number of spells per driver in a given time and let θ be the mean number of accidents per spell. Now consider two combined Poisson distributions with parameters λ and θ. Then from first principles the probability of man having no accidents is given by

$$\begin{aligned} P(0) &= e^{-\lambda}+e^{-\lambda}\frac{\lambda}{1!}e^{-\theta}+e^{-\lambda}\frac{\lambda^2}{2!}e^{-2\theta}+\dots \\ &= e^{-\lambda}\left(1+\frac{\lambda e^{-\theta}}{1!}+\frac{(\lambda e^{-\theta})^2}{2!}+\dots\right) \\ &= \exp[\lambda(e^{-\theta}-1)]. \end{aligned}$$

Since the generating function (g.f.) for the Poisson distribution with parameter θ is $e^{\theta(s-1)}$, the g.f. for the number of accidents in k spells is $e^{k\theta(s-1)}$ and so the g.f. for the final accident distribution is

$$\sum \frac{e^{-\lambda}\lambda^k}{k!}[e^{k\theta(s-1)}] = \exp[-\lambda+\lambda e^{\theta(s-1)}]$$

$$= \exp[\lambda(e^{\theta(s-1)}-1)], \qquad (10.3)$$

and the moment generating function (m.g.f.) is

$$\exp[\lambda(e^{\theta(e^t-1)}-1)]. \qquad (10.4)$$

Equation (10.3) gives the accident probabilities $P(r)$ immediately. For instance

$$P(0) = \exp[\lambda(e^{-\theta}-1)]$$

$$P(1) = \frac{d}{ds}\{\exp[\lambda(e^{\theta\, s-1)}-1)]\}|_{s=0}$$

$$= \lambda e^{-\theta}.P(0)$$

$$P(2) = \frac{d^2}{ds^2}\{\exp[\lambda(e^{\theta(s-1)}-1)]\}|_{s=0}$$

$$= \frac{\theta^2\lambda e^{-\theta}}{2!}\{1+\lambda e^{-\theta}\}.P(0)$$

etc.

In general

$$\exp[\lambda(e^{\theta(s-1)}-1)] = \exp[-\lambda(1-e^{-\theta})+\lambda e^{-\theta}(e^{\theta s}-1)].$$

Also
$$(e^{\theta s}-1)^k = \sum_{r=0}^{\infty}\frac{\Delta^k(0^r)}{k!}\theta^r s^r,$$

where $\Delta^k(0^r)$ is the coefficient of $\theta^r/r!$ in the expansion of $(e^\theta-1)^k$ for $k>0$ and $0 \leqslant r \leqslant k-1$. This quantity is otherwise known as the kth difference of the rth power of zero. An exposition of this topic can be found in *The Calculus of Observations* by Whittaker and Robinson (1924). A table of values $\Delta^k(0^r)/k!$, $n=2(1)25$, $k=2(1)r$ can be found in *Statistical Tables for Agricultural, Biological and Medical Research* (1953).

Using the above result we have

$$\exp[\lambda e^{-\theta}(e^{\theta s}-1)] = \sum_{r=0}^{\infty}\frac{s^r\theta^r}{r!}\sum_{k=0}^{r}\frac{\Delta^k(0^r)}{k!}(\lambda e^{-\theta})^k.$$

This distribution was called the 'Long' distribution by Cresswell and Froggatt, but it can be shown to be formally equivalent to the Neyman Type A. It can easily be shown directly or by means of equation (10.4) that the mean of this distribution is $\lambda\theta$ and the variance $\lambda\theta(1+\theta)$. If we have a reasonably large sample we can set the sample mean and variance equal to these quantities in order to determine the parameters λ and θ, and hence obtain the probability (or frequency) distribution.

We have assumed in this section that accidents happen only in a spell, i.e. we have implied that culpable accidents alone are considered. If we relax this assumption whilst retaining the other, and assume in addition that accidents can occur outside a spell to all drivers, but that these are rare and occur independently both of spells and of accidents within a spell, we can deduce yet another probability distribution, called the 'Short' distribution. Let ϕ denote the mean number of these additional accidents, then the generating function for the resultant distribution is

$$\exp\left[\lambda(e^{\theta(s-1)}-1)+\phi(s-1)\right].$$

It can then be deduced that for $r \geqslant 1$ the probabilities are given by

$$P(r) = P(0)\sum_{j=0}^{r}\left\{\frac{\theta^{r-j}}{(r-j)!}\frac{\phi^{j}}{j!}\sum_{k=0}^{j}\left[\frac{\Delta^{k}(0^{j})}{k!}(\lambda e^{-\theta})^{k}\right]\right\},$$

where $P(0)=\exp\left[\lambda(e^{-\theta}-1)-\phi\right]$.
For example,

$$P(1) = \exp\left[\lambda(e^{-\theta}-1)-\phi\right]\left[\frac{\phi}{1!}+\frac{\theta}{1!}\frac{(\lambda e^{-\theta})}{1!}\right].$$

Then λ, θ, and ϕ can be obtained from the equalities

$$\begin{aligned}\mu_1' &= \lambda\theta+\phi,\\ \mu_2 &= \lambda\theta(1+\theta)+\phi,\\ \mu_3 &= \lambda\theta(1+3\theta+\theta^2)+\phi.\end{aligned}$$

Note, however, that for small samples the last equation for μ_3 may be unreliable: the relation

$$\phi = q\mu_1',$$

where q is the assumed proportion of chance accidents to all accidents, can be used instead. This produces alternative solutions for θ and λ as follows:

$$\theta = \frac{\mu_2 - \mu_1'}{\mu_1'(1-q)},$$

$$\lambda = \frac{\mu_1'(1-q)}{\theta}.$$

Both the Long and Short distributions are based on more sophisticated models than the negative binomial. Unfortunately, however, although the point seems to have been missed by the authors, interpretation of these distributions is just as equivocal, as they are both capable of either a 'spell' or a 'proneness' interpretation. Irwin (1964) makes this clear in a review of the book by Cresswell and Froggatt.

The factorial moment generating function corresponding to (10.3) is, replacing $(s-1)$ by t,

$$\exp[\lambda(e^{\theta t}-1)].$$

This could be regarded as the m.g.f. of a 'proneness' distribution which therefore has the generating function

$$\exp[\lambda(s^{\theta}-1)].$$

Hence the same distribution would arise if the individuals were thought of as having 'pronenesses' θ, 2θ, 3θ, ... with

$$p(r\theta) = \frac{e^{-\lambda}\lambda^r}{r!}.$$

The Short distribution, in a similar way, could be interpreted as adding a general random risk to the 'proneness' just mentioned.

In theory, the corresponding bivariate frequency distributions should afford a means of discriminating between the two hypotheses, although the actual forms of the distributions may well turn out to be intractable in practice. One point to notice however, is that if the spell hypothesis is the right one the distributions in two different exposure periods should be completely independent, i.e. the correlation coefficients between the accidents sustained by the same individuals in the two periods should be zero. This is a consequence of the fact that all individuals start alike, that spells occur randomly with one Poisson distribution and accidents within spells with another, the parameters being the same for all individuals.

The data of Cresswell and Froggatt in fact show small but significant positive correlations, and these have been reported elsewhere in the literature. Irwin has pointed out that these are difficult to interpret. One explanation for them could be inequality of exposure to risk or of the tendency to report accidents. However, they do tend to contradict the spell hypothesis.

V *Distributions of time-intervals between successive accidents*

Study of the time-intervals between successive accidents can also be of some use in accident analysis.

Suppose that we have a sample of drivers with equal exposure times of x years. Then a driver who incurs at least one interval between two accidents of less than 1 year, say, will be called, after Cresswell and Froggatt, a 'repeater', and one who incurs at least two such intervals a 'double repeater'.

Now the interval distribution, conditional on k accidents in the whole period is the same on both the random and proneness hypotheses, and so also are the expected numbers of repeaters. Consideration of these number does not, therefore, afford a means for distinguishing between the two hypotheses.

If a line of length unity is divided randomly into n intervals by $(n-1)$ points, including the end points, the probability that just i of the intervals exceed g has been given by Irwin (1955) as

$$P(i > g) = \frac{n!}{i!(n-i)!}\left\{(1-ig)^{n-1} - \binom{n-i}{1}[1-(i+1)g]^{n-1} + \dots \right.$$
$$\left. + (-1)^j \binom{n-i}{j}[1-(i+j)g]^{n-1} + \dots \right\}, \qquad (10.5)$$

the series continuing as long as $[1-(i+j)g] > 0$.

For x years exposure and an interval of 1 year, the value of g is $1/x$. This is not quite what is required because we do not wish to include the end-intervals. If, however, $\sum y$ years is the sum of the $(k-1)$ intervals between the k accidents, and g is set equal to $1/\sum y$, equation (10.5) still holds with $n=(k-1)$. It gives the exact probability conditional on $\sum y$, and is true for either hypothesis. Putting $i=1,2$ in $[1-P(i>g)]$ gives the probabilities of repeaters and double repeaters. In practice only the first term or terms may be required.

However it is possible to approximate to the probabilities that an individual who has k accidents will be a repeater or double repeater, and hence to obtain the expected number of repeaters for each value of k.

In the random case we make use of the fact that the distribution of intervals between accidents is negative exponential. If accidents occur at random on the average once in μ units, the average number of accidents in time t is t/μ and the probability of one or more accidents in $(0, t)$ is

$$1-\exp(-t/\mu).$$

Now consider an individual having k accidents in x years. Then by the above result, writing k/x for the true mean interval between accidents, the approximate probability of any particular accident not being followed by another within 1 year, say, is

$$q = \exp(-k/x)$$

and

$$p = 1-\exp(-k/x).$$

But the individual had k accidents, and hence $(k-1)$ time-intervals between consecutive accidents. Thus the probability that none of the $(k-1)$ time-intervals is less than 1 year is

$$q^{k-1} = \exp[-k(k-1)/x],$$

and hence the probability of at least one interval less than 1 year is

$$P(\geqslant 1, k) = 1-q^{k-1} = 1-\exp[-k(k-1)/x].$$

This then is a 'repeater'.

If N men have incurred k accidents each, then

$$E(N) = N\{1-\exp[-k(k-1)/x]\}.$$

If for example 4 men have 7 accidents in 14 years,

$$N = 4, \quad k = 7, \quad x = 14$$

and

$$\begin{aligned} E(N) &= 4[1-\exp(-7\times 6/14)] \\ &= 4(1-\exp(-3)] \\ &\approx 3{\cdot}7. \end{aligned}$$

This idea can be extended to further accidents, three for example. Suppose N men have k accidents in x years, then the distribution of repeated time-intervals of less than a year is given approximately by the binomial $(q+p)^{k-1}$. In particular, the probability of no time intervals less than a year is given by

$$P(0,k) = q^{k-1},$$

and the probability of one such interval is

$$P(1,k) = (k-1)pq^{k-2}.$$

Thus the probability of a man incurring less than 2 time-intervals less than 1 year is

$$\begin{aligned} P(<2,k) &= P(0,k)+P(1,k) \\ &= q^{k-1}+(k-1)pq^{k-2} \\ &= (k-1)q^{k-2}-(k-2)q^{k-1} \\ &= (k-1)\exp[-k(k-2)/x]-(k-2)\exp[-k(k-1)/x], \end{aligned}$$

and then

$$P(\geqslant 2,k) = 1-P(<2,k)$$

and this can be called a 'double repeater'. For N men, $E(N) = N \times P(\geqslant 2,k)$. If, for example, $N=10$, $k=5$ and $x=15$ we have

$$q = \exp(-1/3) = 0{\cdot}7166, \quad p = 0{\cdot}2834$$

then $$P(1,5)+P(0,5) = 4pq^2+q^4$$
$$= 0{\cdot}681$$

and $$E(N) = 10[1-0{\cdot}681]$$
$$\approx 3{\cdot}2.$$

Let us now consider the 'proneness' negative binomial case. If accident proneness exists in the population the following approximate argument would give the time between accidents.

If the liability of a man to accident over a long period is λ, the probability of k accidents is

$$\frac{e^{-\lambda}\lambda^k}{k!}.$$

If liability is distributed in a gamma distribution, i.e. the proportion with liability λ is

$$P(\lambda) = \frac{c^p}{\Gamma(p)} e^{-c\lambda}\lambda^{p-1},$$

the combined probability of a man being subject to liability λ and also having k accidents is given, as in subsection I of this section, as

$$P(\lambda, k) = \frac{c^p}{k!\Gamma(p)} e^{-(c+1)}\lambda^{p+k-1}.$$

Integrating with respect to λ, $0 \leqslant \lambda < \infty$, gives the probability of a man selected at random having k accidents as

$$P(k) = \left(\frac{c}{c+1}\right)^p \frac{\Gamma(p+k)}{k!\Gamma(p)(c+1)^k},$$

which is a negative binomial with parameters c and p. Now the probability that a man who incurred k accidents has liability λ is given by the rule for conditional probability as

$$\begin{aligned} P(\lambda|k) &= P(\lambda, k)/P(k) \\ &= \frac{(c+1)^{p+k}}{\Gamma(p+k)} . e^{-(c+1)\lambda}\lambda^{p+k-1}. \end{aligned}$$

We now need an estimate of the individual's true λ. Cresswell and Froggatt take the mode $(p+k-1)/(c+1)$, but this estimate is biased on the low side. The mean $(p+k)/(c+1)$ would be better. Better still, since the probability of a non-repeater is taken as $\exp[-(k-1)\lambda/x]$ would be to average this expression over the λ distribution, which leads to

$$\bar{\lambda} = [1+(k-1)/(c+1)]^{-(p+k)}.$$

For an individual having k accidents in x years an accident would most probably occur on the average (i.e. over a long time) once every m years where

$$m = x/\bar{\lambda}.$$

Thus the probability of an accident not occurring in less than 1 year is given by

$$q = \exp[-\bar{\lambda}/x].$$

This applies to x consecutive years chosen at random. Consider N men having k accidents in x years. Now k accidents imply $(k-1)$ time-intervals, and the probability of $(k-1)$ intervals less than 1 year is

$$q^{k-1} = e^{-\xi},$$

where $$\xi = \frac{(k-1)\bar{\lambda}}{x}.$$

Hence the probability of a repeater is

$$1-q^{k-1} = 1-\mathrm{e}^{-\xi},$$

and $$E(N) \quad = N(1-\mathrm{e})^{-\xi}.$$

A similar formula holds for double repeaters.

Note that the expressions obtained by these approximate methods should agree for both hypotheses: the exact expression is the same for both. Since this is available, and should not require an undue amount of calculation, there seems little point in using an approximation.

Although these results are useless for the purpose of discrimination between the random and accident-proneness hypotheses, time-interval distributions can be used for this purpose. For an individual who has k accidents the probability that an interval is less than y years is

$$(1-g)^{k-1},$$

where $$g = y/\sum y,$$

$$\sum y = \text{sum of intervals between accidents.}$$

If $P(k)$ is the probability that an individual has k accidents, the total probability for all individuals is

$$\sum P(k)(1-g).^{1-k}$$

The interval distributions obtained by equating the $P(k)$ either to the observed, expected Poisson, or expected negative binomial frequencies, could thus be obtained. For most accident data we already know that the negative binomial is a better fit than the Poisson so that this kind of analysis could really only corroborate the fact.

We end with a general comment about the study of accident data. These studies are useful only if, in the end, they lead to a means of identifying accident-prone individuals. It seems unlikely that these can be identified by purely statistical means, and recourse may have to be had to others. One such method which has been tried is to use psychological tests in an attempt to find some personal characteristic which is correlated with liability to accident. The results of experiments on these lines are very variable and largely disappointing.

Obvious possibilities of a simple kind such as speed of reaction or good eyesight have turned out to be of no help: the situation is evidently too complex for a simple relationship of this kind. The most promising means of detecting accident-prone individuals may involve considering temperament and accident record together. If this is done, however, injustice may be done to some individuals, and although this may be justified in the case of professional drivers in transport organizations, it could not be applied generally.

References

Chapter 1

ALMOND, J. (*ed.*) (1965). *Proceedings of the Second International Symposium on the Theory of Traffic Flow*, London 1963. Paris: OECD.

BUCHANAN, C. (1963). *Traffic in Towns*. London: H.M. Stationery Office.

GREENSHIELDS, B. D. (1935). 'A study of traffic capacity.' *Proc. High Res. Bd.*, **14**, 448–474.

HERMAN, R. (*ed.*) (1961). *Theory of Traffic Flow*. Amsterdam: Elsevier.

LIGHTHILL, M. J. and WHITHAM, G. B. (1955). 'On kinematic waves II. A theory of traffic on long crowded roads.' *Proc. Roy. Soc.* A, **229**, 317–345.

Chapter 2

EDIE, L. C. (1965). 'Traffic stream measurements and definitions.' *Proc. Sec. Internat. Symposium on the Theory of Traffic Flow*, London, 1963. Paris: OECD.

JONES, T. R. and POTTS, R. B. (1962). 'The measurement of acceleration noise – a traffic parameter.' *Opns. Res.*, **10**, 745–763.

HELLY, W. and BAKER, P. G. (1965). 'Acceleration noise in a congested signalized environment.' *Proc. Third Internat. Symposium on the Theory of Traffic Flow*, New York, 1965.

LIGHTHILL, M. J. and WHITHAM, G. B. (1955). 'On kinematic waves II. A theory of traffic on long crowded roads.' *Proc. Roy. Soc.* A, **229**, 317–345.

WARDROP, J. G. (1952). 'Some theoretical aspects of road traffic research.' Road Paper 36, *Proc. Instn. Civ. Engnrs.* Pt. 2, **1**, 325–378.

Chapter 3

CARSLAW, H. S. and JAEGER, J. C. (1941). *Operational Methods in Applied Mathematics*. London: O.U.P.

EDIE. L. C. (1961). 'Car-following and steady-state theory for non-congested traffic.' *Opns. Res.*, **9**, 66–75.

HERMAN, R., MONTROLL, E. W., POTTS, R. B. and ROTHERY, R. W. (1959). 'Traffic dynamics: analysis of stability in car-following.' *Opns. Res.*, **7**, 86–106.

HERMAN, R. and POTTS, R. B. (1961). 'Single-lane traffic theory and experiment.' *Theory of Traffic Flow*. Amsterdam: Elsevier.

KOMETANI, E. and SASAKI, T. (1958). 'On the stability of traffic flow I.' *J. Opns. Res. Japan*, **2**, 11–26.

KOMETANI, E. and SASAKI, T. (1961). 'Dynamic behaviour of traffic with a non-linear spacing-speed relationship.' *Theory of Traffic Flow.* Amsterdam: Elsevier.

LIGHTHILL. M. J. and WHITHAM, G. B. (1955). 'On kinematic waves II. A theory of traffic on long crowded roads.' *Proc. Roy. Soc.* A, **229**, 317–345.

NEWELL, G. F. (1962a). 'Non-linear effects in the dynamics of car-following.' *Opns. Res.*, **9**, 209–229.

NEWELL, G. F. (1962b). 'Theories of instability in dense highway traffic.' *J. Opns. Res. Japan*, **5**, 9–54.

NEWELL, G. F. (1965). 'Theories of instability in dense highway traffic.' *Proc. Sec. Internat. Symposium on the Theory of Traffic Flow*, London, 1963. Paris: OECD.

SASAKI, T. (1959). 'On the stability of traffic flow II.' *J. Opns. Res. Japan*, **2**, 60–79.

Chapter 4

ANDERSON, R., HERMAN, R. and PRIGOGINE, I. (1962). 'On the statistical distribution function theory of traffic flow.' *Opns. Res.*, **10**, 180–196.

FRANKLIN, R. E. (1962). 'The propagation of disturbances in a single lane of traffic.' *Civ. Eng. Rev.*, **57**, 344–346, 487–489.

LIGHTHILL, M. J. and WHITHAM, G. B. (1955). 'On kinematic waves II. A theory of traffic on long crowded roads.' *Proc. Roy. Soc.* A, **229**, 317–345.

NEWELL, G. F. (1962). 'Theories of instability in dense highway traffic.' *J. Opns. Res., Japan*, **5**, 9–54.

PRIGOGINE, I., HERMAN, R. and ANDERSON, R. (1965). 'Further developments in the Boltzmann-like theory of traffic flow.' *Proc. Sec. Internat. Symposium on the Theory of Traffic Flow,* London, 1963. Paris: OECD.

Chapter 5

GREENBERG, H. (1959). 'An analysis of traffic flow.' *Opns. Res.*, **7**, 79–85.

GREENSHIELDS, B. D. (1935). 'A study of traffic capacity.' *Proc. High. Res. Bd.*, **14**, 448–474.

HAIGHT, F. A. (1958). 'Towards a unified theory for road traffic.' *Opns. Res.*, **6**, 813–826.

Chapter 6

BECKMAN, M., MCGUIRE, C. B. and WINSTEN, C. B. (1956). '*Studies in the Economics of Transportation.* New York: Yale University Press.

BUCHANAN, C. (1963). *Traffic in Towns.* London: H.M. Stationery Office.

CHARLESWORTH, G. and WEBSTER, F. V. (1958). 'Some factors affecting the capacity of intersections controlled by traffic signals.' *4th Int. Study Wk. in Tr. Engng*, 1958, Copenhagen, Denmark. London: World Touring and Automobile Organisation.

HAIGHT, F. A. (1959). 'Overflow at a traffic light.' *Biometrika*, **46**, 420–424.

TANNER, J. C. (1962). 'A theoretical analysis of delays at an uncontrolled intersection.' *Biometrika*, **49**, 163–170.

UEMATU, T. (1958). 'On traffic control at an intersection controlled by the repeated fixed-cycle traffic light.' *Ann. Inst. Statist. Math.* (*Tokyo*), **9**, 87–107.

WEBSTER, F. V. (1958). 'Traffic signal settings.' Technical Paper 39, Road Research Laboratory D.S.I.R.

WEBSTER, F. V. and WARDROP, J. G. (1962). 'Capacity of urban intersections.' *Traff. Engng & Control*, **4**, 396–401.

Chapter 7

BOREL, E. (1942). 'Sur l'emploi du théorème de Bernouilli pour faciliter le calcul d'un infinité de coefficients. Application au problème de l'attente a un guichet.' *Comptes Rendus Acad. Sci. Paris*, **214**, 452–456.

COX, D. R. and SMITH, W. L. (1961). *Queues*. London: Methuen. New York: Wiley.

MILLER, A. J. (1961a). 'A queueing model for road traffic flow.' *J. R. Statist Soc.* B, **23**, 64–75.

MILLER, A. J. (1961b). 'Traffic flow treated as a stochastic process.' *Theory of Traffic Flow*. Amsterdam: Elsevier.

MILLER, A. J. (1962). 'Road traffic flow considered as a stochastic process.' *Proc. Camb. Phil. Soc.*, **58**, 312–325.

RALLIS, T. 'The application of Erlarg's theory to the calculation of road capacity.' *Proc. Third Internat. Symposium on the Theory of Traffic Flow*, New York, 1965.

SMEED, R. J. and BENNET, G. T. (1949). 'Research on road safety and traffic flow.' Road Paper 29, *Proc. Instn. Civ. Engnrs.*

TANNER, J. C. (1953). 'A problem of interference between two queues.' *Biometrika*, **40**, 58–69.

Chapter 8

ADAMS, W. F. (1936). 'Road traffic considered as a random series.' *J. Instn. Civ. Engnrs.*, **4**, 121–130.

BECKMAN, M., MCGUIRE, C. B. and WINSTEN, C. B. (1956). *Studies in the Economics of Transportation*. New York: Yale University Press.

BLUNDEN, W. R. and PRETTY, R. L. (1965). 'On the theory of deterministic cyclic traffic flows in networks.' *Proc. Third Internat. Symposium on the Theory of Traffic Flow*, New York, 1965

CLAYTON, A. J. H. (1941). 'Road traffic calculations.' *J. Instn. Civ. Engnrs.*, **16**, 247–284.

DARROCH, J. N. (1964). 'On the traffic light queue.' *Ann. Math. Statist.*, **35**, 380–388.

GARWOOD, F. (1940). 'An application of the theory of probability to the operation of vehicular-controlled traffic signals.' *J. R. Statist. Soc.* B, **7**, 65–77.

HAIGHT, F. A. (1963). *Mathematical Models for Traffic Flow.* New York: Academic Press.

HAWKES, A. G. (1965). 'Queueing for gaps in traffic.' *Biometrika,* **52,** 1–6.

JEWELL, W. S. (1961). 'Waiting for a gap in traffic.' Research Report 6, Opns. Res. Centre, University of California, Berkeley.

JEWELL, W. S. (1962). 'Multiple entries in traffic.' Research Report 29, Opns. Res. Centre, University of California, Berkeley.

KENDALL, D. G. (1951). 'Some problems in the theory of queues.' *J. R. Statist. Soc.* B, **13**, 151–185.

LITTLE, J. D. C. (1961). 'Approximate expected delays for several manoevres by a driver in Poisson traffic.' *Opns. Res.*, **9**, 39–52.

MILLER, A. J. (1963). 'Analysis of bunching in rural two-lane traffic.' *Opns. Res.*, **11**, 236–247.

NEWELL, G. F. (1960). 'Queues for fixed-cycle traffic light.' *Ann. Math. Statist.*, **31**, 589–597.

NEWELL, G. F. (1965). 'Approximation methods for queues with application to the fixed-cycle traffic light'. *SIAM Review* **7**, 223–239

OLIVER, R. M. (1962). 'Distribution of gaps and blocks in a traffic stream.' *Opns. Res.*, **10**, 197–217.

OLIVER, R. M. and BISBEE, E. F. (1962). 'Queueing for gaps in high flow traffic streams.' *J. Opns. Res.*, **10**, 105–114.

RAFF, M. S. (1951). 'Distribution of blocks in an uncongested stream of automobile traffic.' *J. Am. Statist. Assoc.*, **46**, 114–123.

TANNER, J. C. (1951). 'The delay to pedestrians crossing a road.' *Biometrika*, **38**, 383–392.

TANNER, J. C. (1953). 'A problem of interference between two queues.' *Biometrika*, **40**, 58–69.

TANNER, J. C. (1961). 'Delays on a two-lane road.' *J. R. Statist. Soc.*, **23**, 38–63.

TANNER, J. C. (1962). 'A theoretical analysis of delays at an uncontrolled intersection.' *Biometrika*, **49**, 163–170.

WEISS, G. H. and MARADUDIN, A. A. (1962). 'Some problems in traffic delay.' *J. Opns. Res.*, **10**, 74–104.

WELDING, P. I. (1965). 'Time series analysis as applied to traffic flow.' *Proc. Sec. Internat. Symposium on The Theory of Traffic Flow*, London, 1963. Paris: OECD.

YEO, G. F. (1964). 'Traffic delays on a two-lane road.' *Biometrika*, **51**, 11–16.

Chapter 9

BENSON, J. C. (1962). 'A program to simulate on a Pegasus II computer the behaviour of traffic at a single intersection controlled by vehicle-actuated traffic signals.' Road Research Laboratory D.S.I.R. Note No. LN/182/JCB. Unpublished.

BUTCHER, J. C. (1961). 'Random sampling from the Normal distribution.' *Computer J.*, **3**, 251–253.

EVANS, D. H., HERMAN, R. and WEISS, G. H. (1964). 'The highway merging and queueing problem.' *Opns. Res.*, **12**, 832–857.

EVANS, D. H., HERMAN, R. and WEISS, G. H. (1965). 'Queueing at a stop sign.' *Proc. Sec. Internat. Symposium on the Theory of Traffic Flow*, London, 1963. Paris: OECD.

GERLOUGH, D. L. (1957). 'Simulation of freeway traffic by digital computers.' *Proc. Conf. on Increasing High. Engn. Productivity*, Georgia Institute of Technology, 1956.

GOODE, H. H., POLLMAR, C. H. and WRIGHT, J. B. (1956). 'The use of a digital computer to model a signalised intersection.' *Proc. High. Res. Bd.*, **35**, 548–557.

MATHEWSON, J. H., TRAUTMAN, D. L. and GERLOUGH, D. L. (1955). 'Study of traffic flow by simulation.' *Proc. High. Res. Bd.*, **34**, ,522–530.

TOCHER, K. D. (1963). *The Art of Simulation*. London: O.U.P.

Chapter 10

ARBOUS, A. G. and KERRICH, J. C. (1951). 'Accident statistics and the concept of accident proneness.' *Biometrics*, **7**, 340–432.

CRESSWELL, W. L. and FROGGATT, P. (1963). *The Causation of Bus Driver Accidents: an Epidemiological Study*. London: O.U.P.

FARMER, E. and CHAMBERS, E. G. (1926). 'A psychological study of individual differences in accident rates.' Rep. Industr. Fat. Res. Bd., No. 38.

FISHER, R. A. and YATES, F. (1953). Table XXII, *Statistical Tables for Agricultural, Biological and Medical Research*. London: Oliver and Boyd.

GREENWOOD, M. and WOODS, H. M. (1919). 'A report on the incidence of industrial accidents upon individuals with special reference to multiple accidents.' Rep. Industr. Fat. Res. Bd., No. 4.

GREENWOOD, M. and YULE, G. U. (1920). 'An enquiry into the nature of frequency distributions representative of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents.' *J. R. Statist. Soc.*, **83**, 255–279.

HÄKKINEN, S. (1958). 'Traffic Accidents and Driver Characteristics. A Statistical and Psychological Study.' Helsinki: Finland's Institute of Technology Scientific Researches, No. 13.

IRWIN, J. O. (1955). 'A unified derivation of some well-known frequency distributions of interest in biometry and statistics.' *J. R. Statist. Soc.* A, **118**, 389–398.

IRWIN, J. O. (1964). 'The personal factor in accidents.' *J. R. Statist. Soc.* A, *127* 438–451.

NEWBOLD, E. M. (1926). 'Contributions to the study of the human factor in the causation of accidents.' Rep. Industr. Hlth. Res. Bd. Lond., No. 34.

NEWBOLD, E. M. (1927). 'Practical applications of the statistics of repeated events, particularly in industrial accidents.' *J. R. Statist. Soc.* **90**, 487–547.

MCKENDRICK, A. G. (1914). 'Studies on the theory of continuous probabilities with special reference to its bearing on natural phenomena of a progressive nature.' *Proc. Lond. Math. Soc.* (2), **13**, 401–416.

MCKENDRICK, A. G. (1926). 'Applications of mathematics to medical problems.' *Proc. Edin. Math. Soc.*, **44**, 98–130.

STUDENT (1907). 'On the error of counting with a haemacytometer.' *Biometrika*, **5**, 351–360.

WHITTAKER, E. T. and ROBINSON, G. (1924). *The Calculus of Observations*. London: Blackie & Son.

Subject Index

Acceleration noise, 16–17, 38
Accident proneness, 149–50, 157, 165, 167; *see also* Negative binomial, 'Long' and 'Short' distributions
Accident statistics, 148–52; *see also* Negative binomial, Bivariate distributions
Area traffic control, 72–6

Before-and-after studies, 3, 69
Beta distribution, 54, 87.
Bottlenecks, 133
Bivariate distributions, 157–59, 162

Capacity, 18, 63
  design, 19
  possible, 18
  practical, 19
Car-following, *see* Follow-the-leader theory
Computers, 5, 73–5, 135–7
Concentration, 7–9
Continuum theory, *see* Kinematic theory

Delay
  at traffic signals, 59–61
  on an open road, 132
  pedestrian, 83, 110–13
  static, 106; *see also* Priority intersections, Stochastic models
Density, *see* Concentration

Flow, 7–9
  steady-state, 24, 28–35; *see also* Flow-concentration relation
Flow-concentration relation, 8, 17–19, 32, 41–2; *see also* Fundamental diagram
Fluid analogy, *see* Kinematic theory
Follow-the-leader theory, 4, 21–38, 47
Fundamental diagram, 17–20, 52–6

Gamma distribution, 88–9, 153, 165
Gap-acceptance, 106–9, 140

Intersections
  priority, 113–24; *see also* Static delay
  signal-controlled, 59, 124–31; *see also* Traffic signals
  simulation of, 140–3

Kinematic theory, 4, 22, 37, 39–47, 133
Kinetic theory, 48–51

'Long' distribution, 159–60

Mean speed
  space, 9–11
  time, 9–11

Negative binomial, 152—5, 165
Negative exponential, 78, 81–5, 95–6, 145
  delayed, 83, 108

Origin-destination survey, 3
Overtaking models, 102–5

Pearson's distributions, 87–8
Pedestrian control experiment, 67–70
Pedestrian-crossings, 67, 69

Poisson distribution, 77–80, 145, 151–2
  generalized, 89

Queueing theory, 90–102
  applications of, 99, 102–5
  non-equilibrium, 100–2
Queue-length distribution, 93–4, 97, 99, 141–3

Random distribution, 77–87; *see also* Poisson distribution and Negative exponential

Sampling from a distribution, 143–7
Sensitivity, 23–4, 36
'Short' distribution, 161–2
Simulation, 5, 135–40; *see also* Intersections
Stability, 24
  asymptotic, 28
  local, 25–8
Stimulus-response equation, 23–4, 35
Stochastic models, 5, 106–31

Time-interval distribution, 163–7
Time series, 3, 133–4
T-junction, *see* Priority intersections
Traffic intensity, 93–4, 99–102
Traffic signals, 57,
  fixed-cycle, 57, 59, 63
  optimum setting of, 63–6
  vehicle-actuated, 58, 62;
  see also Signal-controlled intersections

Waiting-time distribution, 94–8
Waves
  dynamic, 40
  kinematic, 40–2, 45
  shock, 44, 46
  velocity of, 41–2, 46–7

Author Index

ADAMS, W. F., 108, 171
ALMOND, J., 169
ANDERSON, R., 48, 51, 170
ARBOUS, A. G., 151, 173

BAKER, P. G., 17, 169
BECKMAN, M., 62, 108, 170, 171
BENNET, G. T., 171
BENSON, J. C., 138, 172
BISBEE, E. F., 108
BLUNDEN, W. R., 126, 171
BOREL, E., 85, 86, 87, 171
BUCHANAN, C., 2, 76, 169, 170
BUTCHER, J. C., vii, 145, 173

CAPELLE, D. G., vii, 25
CARSLAW, H. S., 25, 169
CHAMBERS, E. G., 150, 173
CHARLESWORTH, G., 60, 171
CLAYTON, A. J. H., 126, 129, 171
COX, D. R., viii, 91, 102, 171
CRESSWELL, W. L., 149, 151, 159, 161, 162, 163, 166, 173.

DARROCH, J. N. 126, 171

EDIE, L. C., vii, 14, 23, 169
ERLANG, A. K., 90, 104
EVANS, D. H. 140, 173

FARMER, E., 150, 173
FISHER, R. A., 173
FRANKLIN, R. E., 47, 170
FROGGATT, P., 149, 151, 159, 161, 162, 163, 166, 173

GARWOOD, F., viii, 116, 117, 172
GERLOUGH, D. L., vii, 137, 138, 173
GOODE, H. H., 137, 173
GREENBERG, H., 170
GREENSHIELDS, B. D., 3, 169, 170
GREENWOOD, M., 148, 151, 152, 154, 173

HAIGHT, F. A., vii, 53, 55, 63, 110, 170, 172
HÄKKINEN, S. 149, 173.
AHWKES, A. J., viii, 109, 126, 172
HELLY, W., 17, 169
HERMAN, R., 23, 25, 27, 30, 36, 48, 51, 140, 169, 170, 173

IRWIN, J. O., 151, 162, 163, 173

JAEGER, J. C., 25, 169
JEWELL, W. S., 109, 172
JONES, T. R., 16, 169

KENDALL, D. G., 121, 172
KERRICH, J. C., 151, 173
KOMETANI, E., 25, 33, 169, 170

LIGHTHILL, M. J., 4, 11, 18, 38, 40, 46, 133, 169, 170
LITTLE, J. D. C., 124, 172

MARADUDIN, A. A., 109, 122, 172
MATTHEWSON, J. H. 137, 173
MCGUIRE, C. B., 62, 108, 170, 171
MCKENDRICK, A. G., 155, 174
MILLER, A. J., 85, 86, 102, 103, 122, 123, 171, 172
MONTROLL, E. W., 25, 169

NEWBOLD, E. M. 151, 174
NEWELL, G. F., vii, 23, 36, 47, 124, 126, 129, 170, 172

OLIVER, R. M., 108, 172

POLLMAR, C. H., 137, 173
POTTS, R. B., 16, 23, 25, 36, 169
PRETTY, R. L., 126, 171
PRIGOGINE, I., 48, 51, 170

RALLIS, T., 103, 171
RAFF, M. S., vii, 113, 172
ROBINSON, G., 160, 174
ROTHERY, R. W., 25, 169

SASAKI, T., 25, 33, 169, 170
SMEED, R. J., 171
SMITH, W. L. 91, 102, 171
'STUDENT' (W. S. GOSSET), 151, 174

TANNER, J. C., vii, 63, 85, 86, 108, 119, 120, 121, 122, 132, 133, 171, 172
TOCHER, K. D., 145, 173
TRAUTMAN, D. L. 137, 173

UEMATU, T., 63, 66, 171

WARDROP, J. G., vii, 9, 11, 12, 63, 169, 171
WEBSTER, F. V., vii, 59, 60, 61, 63, 131, 171
WEISS, G. H., 109, 123, 140, 172, 173
WELDING, P. I., vii, 133, 172
WHITHAM, G. B., 4, 11, 18, 38, 40, 46, 133, 169, 170
WHITTAKER, E. T., 160, 174
WINSTEN, C. B., 62, 108, 170, 171
WOODS, H. M., 151, 152, 173
WRIGHT, J. B., 137, 173

YATES, F., 173
YEO, G. F., 132, 172
YULE, G. U., 148, 151, 154, 173